March 2024

Through the Looking Glass: Overcoming Algorithm Aversion in Accounting

David E. Watson
University of South Florida

Through the Looking Glass: Overcoming Algorithm Aversion in Accounting

by

David E. Watson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a concentration in Accounting
Lynn Pippenger School of Accountancy
Muma College of Business
University of South Florida

Major Professor: Uday S. Murthy, Ph.D.
Lisa Gaynor, Ph.D.
Mark Taylor, Ph.D.
Sandra Schneider-Wright, Ph.D.

Date of Approval:
March 26, 2024

Keywords: Artificial intelligence; complex estimate; decision aid; reliability; transparency; auditor reliance

Dedication

This dissertation is dedicated to those who embrace the challenge of using their ambition and curiosity to positively impact their community and the world around them.

"Life was so much harder, but potentially so much more fulfilling, when you found the courage to choose."

Brandon Sanderson, *Oathbringer*

Acknowledgments

I acknowledge the financial support from the Rene “Bud” Gaiennie Endowment Fund, the Dr. Henry Efebera Scholarship, the Muma College of Business, and the Lynn Pippenger School of Accountancy.

I am deeply grateful to many, many individuals who have left a significant mark on my life. My parents deserve the first mention for nurturing my passions and for their unwavering support in every endeavor. My brother, my companion in countless life experiences, has been a source of constant learning and growth.

Special thanks to David Wood and Uday Murthy, two of many mentors who inspired me to pursue knowledge relentlessly and to question the conventional.

I also thank Ben Commerford, Sean Dennis, Jennifer Joe, and Jenny Ulla for allowing me to adapt their experimental instrument and Abigail Zhang for her support with Explainable AI

Thank you for the dedication of support staff and educators who have profoundly shaped my academic path, instilling in me a deep appreciation for the world's complexities.

I owe much to the authors and creators whose works have moved me, enriching my journey with emotion, purpose, and the drive for a better future.

My Ph.D. cohort—Ahmed, Chandler, and Rich—shared in this unique journey; together, we made it.

Above all, my heartfelt gratitude goes to my wife, whose love, patience, and encouragement have been a constant guide. Her presence has transformed me and prepared us for the next, beautiful chapter of our lives.

Table of Contents

List of Tables ii

List of Figures iii

List of Abbreviations iv

Abstract v

Chapter One: Introduction 1

Chapter Two: Theory and Hypothesis Development 8
- 2.1 Algorithm Aversion and Appreciation 8
- 2.2 Hoff & Bashir Three-Layer Trust in Automation Model 11
- 2.3 Reliability 16
- 2.4 Transparency 20
- 2.5 Interactive Effects of Reliability and Transparency on Reliance 25

Chapter Three: Method 29
- 3.1 Initial Pilot Study 29
- 3.2 Secondary Pilot Study 30
- 3.3 Participants 32
- 3.4 Independent Variables 33
- 3.5 Dependent Variables 33
- 3.6 Experimental Audit Case 36
- 3.7 Experimental Procedure 37

Chapter Four: Results 40
- 4.1 Manipulation and Attention Checks 40
- 4.2 Analysis Methodology 42
- 4.3 Tests of Hypotheses and Research Questions 43
- 4.4 Additional Analyses 56
  - 4.4.1 Parkes' Composite Measure Individual Constructs 56
  - 4.4.2 Source Credibility Theory 57
  - 4.4.3 Mediation and Moderation Process Models 60

Chapter Five: Discussion and Conclusion 67

References 74

Appendix A: Intervention Introduced to Participants 80

Appendix B: IRB Exempt Determination Letter 82

Appendix C: Table 1A – Descriptive of Pre-Scaled Reliance Score...83

Appendix D: Declaration of Use of Generative AI and AI-assisted Technologies 84

List of Tables

Table 1: Demographic Descriptive Statistics 32

Table 2: Hypotheses Tests: Reliance Score on AI Tool Estimate 44

Table 3: Hypotheses Tests: Reliance Score on AI Tool Estimate (Professionals) 46

Table 4: Hypotheses Tests: Reliance Score on AI Tool Estimate (Students) 48

Table 5: Hypotheses Tests: Composite Parkes Measure on AI Tool Estimate 50

Table 6: Hypotheses Tests: Likelihood of Reliance on AI Tool Estimate 53

Table 7: Descriptive Statistics of Post-Experiment Questionnaire on Composite Reliance Measure 56

Table 8: Descriptive Statistics of Post-Experiment Questionnaire 58

Table 9: Mean Difference between AI and Management Bias and Evidence Quality by Conditions 59

Table 1A: Descriptive Statistics of Pre-Scaled Reliance Score 83

List of Figures

Figure 1: Layers of Trust in Automation....12

Figure 2: Factors of Trust in Algorithms....13

Figure 3: Hypothesized Process between Reliability Mechanism and Reliance....19

Figure 4: Hypothesized Simple Effect of Reliability Information....20

Figure 5: Hypothesized Process between Transparency Mechanism and Reliance....25

Figure 6: Hypothesized Simple Effect of Transparency Information....26

Figure 7: Research Questions and Interactive Effects of Factors of Trust....28

Figure 8: Effects of Interventions on Auditor Proposed Estimate....45

Figure 9: Effects of Interventions on Auditor Proposed Estimate – Professionals....47

Figure 10: Effects of Interventions on Auditor Proposed Estimate – Students....49

Figure 11: Effects of Reliability and Transparency on Composite Reliance Measure....51

Figure 12: Effects of Reliability and Transparency on Likelihood of Reliance....54

Figure 13: Serial Mediation Analysis on Composite Reliance Measure....61

Figure 14: Parallel Mediation Analysis on Reliance Score....62

Figure 15: Moderated Mediation Analysis on Reliance Score....64

Figure 16: Mediation Analysis of Difference in Understanding on Reliance Score....65

Abstract

Artificial intelligence (AI) is growing rapidly in the accounting field as audit firms invest heavily in technology to enhance audit efficiency and effectiveness. Existing research reveals two contrasting behaviors: algorithm aversion, where individuals are reluctant to rely on algorithms even if their recommendations are equivalent to humans, and algorithm appreciation, where individuals excessively rely on AI without exercising professional skepticism. This study investigates whether auditors' algorithm aversion can be mitigated through interventions of providing reliability information and transparency information (explainability) about the AI's processes in order to enhance auditor reliance on AI tools. The results indicate that the participants were more inclined to rely on an AI tool when provided with reliability information of the tool's historical performance. However, transparency of the AI process did not significantly increase reliance, likely due to low statistical power. Further analyses suggest confidence in AI and perceived evidence quality produced by AI partially mediate the relation between AI transparency and reliance of the AI tool. These findings suggest that audit firms can enhance auditor reliance on AI tools primarily through providing reliability information.

Chapter One: Introduction

Big 4 accounting firms have invested over $9 billion in artificial intelligence (AI) and other advanced technologies since 2015, including extensive technical training to effectively implement them into audit work (Kapoor, 2020; Sutton et al., 2016). Practitioners have used AI to help identify areas of high risk of material misstatement, help analyze more complete data sets, and automate mundane tasks (AICPA-CIMA, 2022). Raschke et al. (2017) report that auditors have been increasing the use of AI to conduct bank inquiries, routine client matters, and highlight unusual transactions. There have been mixed results in existing research on how individuals respond to advanced technologies like artificial intelligence (AI). Audit research suggests auditors exhibit distrust and are unwilling to rely on AI advice relative to human experts (Commerford et al., 2022). Conversely, organizational behavior and management literature demonstrate that individuals favor algorithmically generated advice over their human counterparts (Logg et al. 2019). Together, the distrust of and unquestioning trust in algorithms is called *algorithm bias*. While evidence suggests that individuals exhibit both aversion to or an appreciation of algorithms, we do not clearly understand the moderating conditions that trigger aversion or the mechanisms that induce algorithm aversion in accounting settings. To focus my investigation, I utilized an experimental study where an audit firm approved AI tool generates a material audit adjustment proposal, provides either reliability information, transparency information, or both reliability and transparency information, and measure the participants' aversion to accepting the proposed audit adjustment.

In this study, I examine the role of moderating conditions and the mechanisms underlying algorithm aversion, specifically through the lens of a conceptual model of trust in automation proposed by Hoff & Bashir (2015). Their model is organized into three layers that specify various constructs explaining when trust in automation does and does not occur. Specifically, the three layers of trust in the Hoff & Bashir (2015) model are dispositional trust, situational trust, and learned trust. Trust in automation contains similar traits to interpersonal trust but has some unique dimensions. Based on a review of 127 studies, Hoff & Bashir (2015) assert that trust can be influenced by an individual's unique traits and disposition, the environment in which the automation exists, and the perceptions of the system's operation. Human factors research focuses mostly on the first two layers of trust (dispositional and situational trust), identifying personality traits, experience, education, external environmental conditions, and complex social processes as important factors affecting trust in those layers (Riedl 2022; Langer and Landers 2021; Makarius et al. 2020).

The third layer of "learned trust" comprises different constructs of what influences humans to develop trust through repeated interactions with algorithms such as AI. Auditors interact with AI in various decision-making settings, and through these repeated interactions, they learn to trust (or not trust) AI. Generally, if these AI tools are deployed by the audit firm itself for use by its auditors in the field, this should be a strong signal of the tool's implied reliability and competence. The audit industry has high levels of regulatory and compliance risk that result in considerable economic consequences for failing to address these risks appropriately. Given these risks, field auditors should reasonably believe that the deployed AI tool is of high quality and relying on it should enhance audit quality. However, recent accounting research indicates that auditors still exhibit aversion to AI tools even when their firm is the one implementing the technology (Commerford et al. 2022). Thus, I specifically investigate potential mechanisms for mitigating auditors' aversion in relying on firm-developed AI tools.

Management research supports Hoff and Bashir's (2015) framework, indicating that among the most important factors that shape human trust in automation is the transparency of the AI and its reliability, which are two elements within the learned trust layer (Glikson and Williams Wooley 2020). In this study, I investigate the effects of these two constructs within the learned trust layer in terms of their influence on auditors' reliance on AI. I define reliability as information regarding the proven consistency of the algorithm based on past (historical) performance. Reliability should influence the auditor's judgment of the AI's competence in completing the task because users can infer the system's likely success rate. I define transparency as an explanation regarding the components, factors, and logic the algorithm uses to operate. Such information, when provided, should demystify the AI implementation, and enhance trust in the decision aid.

Findings from prior information systems literature provide mixed evidence regarding the impact of reliability and transparency on users' reliance on automation. De Vries et al. (2008) provide evidence that individuals tend to trust automation when the system is portrayed as an expert. Koustanai et al. (2012) and Wilkison et al. (2007) provide further evidence that training individuals about a decision aid's reliability improves task performance. These three studies, while not in the accounting or auditing domain, support the notion that increasing reliability saliency of an AI tool should enhance reliance on it. However, Spain (2009) demonstrates that higher quality information made automation errors more salient, increasing distrust of automation. Within the auditing domain, Commerford et al. (2022) provide evidence that auditors inherently distrust AI compared to human counterparts. The apparent distrust or aversion towards AI makes the effect of providing more information about the inner workings of AI or its purported reliability unclear about its influence on auditor behavior. These mixed findings underscore the need for additional research to better understand the factors that lead to increased or decreased reliance on AI. This study focuses on understanding how auditors react to this additional information regarding algorithm transparency and reliability.

My study utilizes a 2x2 between-participants design where I examine the effects of algorithm transparency and algorithm reliability on the extent of auditor reliance on a firm-provided AI tool. I manipulate my first independent variable, the algorithm's reliability, by disclosing (or withholding) positive historical performance of the AI tool. I manipulate my second independent variable, algorithm transparency, by providing a feature-driven (more transparent) or a summary (less transparent) explanation of how the AI tool made its decision. I measure my primary dependent variable, the auditor's reliance on the AI, by capturing the magnitude of a potential audit adjustment recommendation from an AI that participants would accept. Participants were provided with a client generated complex estimate and an AI-tool proposes a material adjustment to the client's figure. The extent to which participants rely on the client's (AI's) estimate would indicate a lower (greater) reliance on the AI. Participants who exhibit lower (higher) reliance on AI would be associated with greater (lower) levels of algorithm aversion. I also employ two alternate dependent measures. The first is the likelihood of reliance on AI (Hodge et al., 2021; Ganbold et al. 2022) and the second is a composite reliance measure adapted from Parkes (2017). The likelihood of reliance measure asks participants how likely they are to rely on AI tools to make similar decisions. The composite reliance measure is based on six different questions that probe into the nuances of how an individual interacts with a decision aid.

Based on the Hoff and Bashir (2015) conceptual model of trust in automation, I predict that algorithm transparency and information about the algorithm's reliability will have individual and interactive effects on auditors' reliance on AI. Hoff and Bashir (2015) speculate that interactions occur between the factors within each layer. However, most of the studies they reviewed investigated the impact of either transparency or reliability information in isolation on algorithm reliance. Increased transparency and understanding of algorithms have inconclusive effects on how individuals identify biased procedures (Wang et al. 2020; Langer et al. 2018; Newman et al. 2020). The inability to appropriately identify biased algorithmic procedures can lead individuals to have a greater distrust of automation or increase "over-trust" of automation. These studies often focus on individual trust factors in automation without studying the potential interactive effects of providing information about reliability and transparency.

I anticipate that providing transparency information will increase participants' general understanding of AI, which will in turn increase reliance on AI. I also expect that providing evidence of the prior reliability of the AI will also increase reliance on the AI. However, I cannot predict whether the increase in reliance will differ between the two interventions. Reliability information should strengthen participants' confidence, and likewise, transparency should increase participants' understanding. Higher levels of confidence and understanding could lead to increased reliance on the AI system. However, there is no strong theoretical basis for predicting how both factors are perceived together. Thus, I state the following research question: Is there a joint effect of reliability information and transparency information on auditors' reliance on an AI tool?

The potential mechanisms underlying how increased transparency and reliability of information influence reliance on AI are important to capture. As potential mediators of the relations between the manipulated factors and reliance on AI, I measure participants' general confidence in AI and their confidence in AI's ability to perform the specific audit task. I further measure the extent to which individuals followed the AI's recommendation, altered their decision process, used the information provided by the AI, and if they followed the recommendations made by the AI tool. Finally, I will measure participants' perceptions of the AI's performance, credibility, and understandability as additional potential mediators.

The study provides evidence that auditors, when provided with reliability information, increase their reliance on the AI tool. However, providing transparency information in terms of the detailed features of the algorithm does not significantly affect their reliance, which may be due to the low statistical power stemming from a small sample size. Process models offer some insight into why the transparency intervention did not influence the participants as hypothesized. Participants' confidence in AI's ability to perform the specific task positively influenced their reliance on the AI tool. Additionally, if participants perceived the quality of management evidence as higher, this belief was associated with a lower reliance on the tool. Conversely, if participants perceived the quality of AI evidence as higher, this belief was associated with greater reliance. When analyzing a mediation model regarding the transparency intervention's relationship with confidence in AI and the moderating impact of the reliability intervention, all of the relationships to the reliance score within this model became insignificant. This finding signifies that that when presented with both the reliability and transparency interventions, the beneficial effects of specific task confidence in AI is weakened.

My study makes three contributions to literature, practice, and regulators. First, my study will contribute to auditing research by improving our understanding of the mechanisms underlying algorithm aversion and by introducing the Hoff and Bashir (2015) framework as the theoretical basis for understanding how algorithm aversion can be mitigated. As accounting firms continue to integrate AI into their processes, an increased understanding of algorithm aversion will help firms design AI aids in a manner that leads to an appropriate level of reliance on them. Accounting researchers are interested in better understanding the mechanisms that can potentially mitigate auditors' bias against algorithms when such bias is unwarranted (Commerford et al., 2022; Zhang et al. 2022). My study also contributes to the information systems literature by adding to a growing stream of studies on algorithm aversion with evidence of the importance of algorithm reliability information and learned trust as mechanisms underlying algorithm aversion.

Second, my study contributes to practice by providing insight into the mechanisms that threaten the effective use of AI tools and provides actionable interventions for firms to reduce such threats. Results from my study provide insight into the potential mechanisms underlying algorithm aversion and inform firms' efforts toward improving auditors' judgment and decision-making processes related to AI. Specifically, the study results can inform audit firms to implement reliability information into their AI-based decision aids, allowing auditors to have more confidence in the ability of the AI system to arrive at its recommendations.

Lastly, my study contributes to regulators because as auditors increasingly use AI in their audits, algorithm aversion could negatively impact audit quality and the reliability of financial information disseminated to capital markets. While transparency-increasing mechanisms seem to not increase reliance in my study, results suggest that reliability information about AI-based aids could help increase trust and confidence in these aids when warranted. Further, an additional understanding of algorithm aversion can further direct regulators to better evaluate the appropriateness of auditors' use of algorithms and emergent technologies.

Chapter Two: Theory and Hypothesis Development

2.1 Algorithm Aversion and Appreciation

Dietvorst et al. (2015, p. 114) define the algorithm aversion phenomenon as situations where "experts and laypeople remain resistant to using algorithms, often opting to use forecasts made by an inferior human rather than forecasts made by a superior algorithm." Participants, across their experiments, preferred human-driven forecasts. The study suggests that two potential components of algorithm aversion are confidence and belief, referring to the degree of trust individuals place in models and the belief that algorithms are needed to make perfect forecasts while not extending the same expectation to humans. These findings from psychology literature tie well into the human factors literature on the perceptions of automation. My study further breaks down the concepts of "confidence" and "belief" by linking reliability information to confidence about the competence of the algorithm and transparency to understanding the algorithm's processes. Investigating how the reliability of the system and the transparency of the AI's processes influence auditors' judgment and decision-making processes should reveal key mechanisms that could help improve auditors' reliance on AI in an increasingly automated world.

Recent accounting research indicates that algorithm aversion strongly influences professionals' judgment and decision-making processes. Commerford et al. (2022) demonstrate that auditors prefer to rely on human advisors compared to AI. Auditors also display more algorithm aversion to subjective (estimate-based) advice than objective (concrete answer based) advice. Despite the same values provided by both the in-house human and the AI advisors, the auditors relied more on management's assertions when presented with a proposed audit adjustment by an AI advisor. Commerford et al. (2022) speculate that auditors viewed human advisors as more credible than AI advisors. The difference is amplified when the human source of information appears more credible than the AI. I extend their study by addressing these speculations by investigating the role of reliability and transparency. My study investigates the effects of informing auditors about the superior relative performance of AI over human counterparts in the past. Additionally, my study examines how auditors' reliance on AI is impacted when detailed information about the AI tool's processes is provided. These two factors can potentially help demystify the black box of AI and provide more confidence to auditors about the AI's capabilities.

Contrarily, studies in the organizational behavior literature provide evidence that individuals may over-rely on algorithms inappropriately. Individuals may not understand how the AI processes information, where the AI's data source comes from, or potential biases in the AI's training that may influence the perceived output (Castelo et al. 2019). Lee & See (2004) find that individuals over-trust the algorithms in a manner that exceeds AI-based systems' capabilities, especially in conditions where the AI is highly reliable. Individuals using highly reliable automation tend to be less likely to detect automation failures (Parasuraman et al. 1997). In a study by Liel and Zalmanson (2020), gig platform workers conformed to erroneous algorithm-generated advice even when the correct task outcome was relatively easy to judge. These tasks were simple perceptual judgment tasks, and the gig workers seemed to rely on the algorithmically generated answers more than crowd-sourced opinions. While auditing is not a simple perceptual task, staff and senior auditors may feel strong external pressure to conform and may be also influenced by budgetary pressures leading them to blindly accept an algorithm's generated evidence.

Logg et al. (2019) acknowledge that while many people exhibit algorithm aversion, individuals exhibit algorithm appreciation in certain conditions and contexts. Multiple experiments provide evidence that individuals preferred the algorithm compared to a human when provided with the same information. Their experiments covered objective situations (predicting weight or age) and subjective situations (predicting song recommendations or romantic attraction). The one case in their study where participants displayed algorithm aversion was academic researchers who attempted to predict the results of the romantic attraction experiment in which an algorithm offered to help the user find a partner. Academics seemed to be skeptical of the algorithm's ability to find a romantic partner and chose to not interact with the algorithm. Logg et al. (2019) suggest that algorithm appreciation is stronger among laypeople but still exists among experts, albeit to a lesser extent. Auditors act as experts within their field, which may decrease the potential effects of algorithm appreciation.

However, the complex judgments and decision-making processes that auditors engage in may lead to an opposite set of behaviors exhibited by the participants in the Logg et al. (2019) study. Specifically, auditors are trained to apply professional skepticism when they analyze audit evidence and tools. AS 1015.07 defines professional skepticism as "an attitude that includes a questioning mind and a critical assessment of audit evidence" (PCAOB Release No. 2022-002). Therefore, auditors who exercise professional skepticism should accept audit evidence only after careful and deliberate scrutiny. Algorithm appreciation is a form of bias suggesting that individuals do not exercise their due diligence or critical assessment of the algorithm, leading to an overreliance on the algorithm beyond its capability.

Existing accounting research tends to support the notion that auditors do not exercise algorithm appreciation but exhibit algorithm aversion (Commerford et al. 2022; Hodge et al. 2021; Ganbold et al. 2022). Algorithm aversion is a form of bias that manifests when individuals discount algorithms and do not rely on properly functioning algorithms when they should. Both researchers and practitioners must understand the mechanisms contributing to algorithm aversion to promote auditors' appropriate use of algorithms to enhance audit efficiency and effectiveness. One of the potential methods of understanding the relationship between auditors and adopting AI technology is examining their trust in AI.

2.2 Hoff & Bashir Three-Layer Trust in Automation Model

Hoff & Bashir (2015) develop a three-layer trust in automation model, demonstrating that trust in automation is directly related to reliance on automation. Their framework is formed from three main trust layers: dispositional, situational, and learned (see Figure 1). Dispositional trust describes an individual's inherent tendency to trust in automation. Situational trust describes the external influences (such as environmental conditions) and internal, context-dependent characteristics (such as experience with technology) of individuals. The third layer is learned trust, which describes how an individual evaluates a system from past or current experiences with technology. Each layer of trust operates independently of the other and contains different components, factors, and influences on the variability of trust. Dispositional trust is difficult to influence or change because personality traits are relatively stable and invariant (van Dongen and van Maanen 2012; Parkes 2017; Makarius et al. 2020).

Prior information systems and management research mostly manipulated the context and traits of the algorithm itself, which pertains to situational trust (Kaplan et al. 2001; Alvarado-

Valencia and Barrero 2014; Langer and Landers 2021). Learned trust allows us to understand the underpinnings of individual thinking without delving into the difficulty of measuring human personalities or changing the algorithm itself. Thus, I focus on the layer of learned trust in investigating potential mechanisms of mitigating algorithm aversion. Figure 1 depicts the relationship between these three layers of trust as explained in Hoff & Bashir's framework.

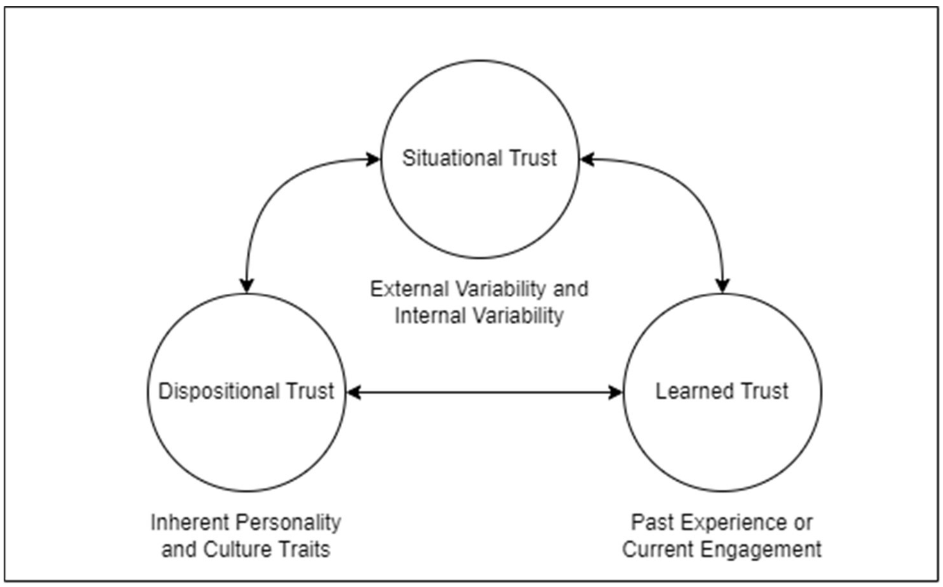

Figure 1. Layers of Trust in Automation
Adapted from Hoff & Bashir (2015)

In Figure 2, my adaptation of their specific framework of the learned trust layer demonstrates the two factors I investigate. Users' pre-existing knowledge impacts their initial reliance strategy before interacting with the algorithm. The difference between learned trust in an individual's experience and learning through external information is that the former is generated through internal experiences, while the latter is generated through external factors. An individual may trust their internal processes more than external information, or, if in a situation where the individual is not confident or familiar with the subject, they may place more weight on external information. Most accountants are not trained in the complex mathematical and computer science methodologies required to comprehend how AI operates, potentially creating an environment where they do not understand the fundamentals of AI, its capabilities, or its limits. The reliability information speaks directly about the system's capabilities, while the transparency information reveals information about how the system works. Both influence a user's reliance strategy and should affect their level of learned trust in the system.

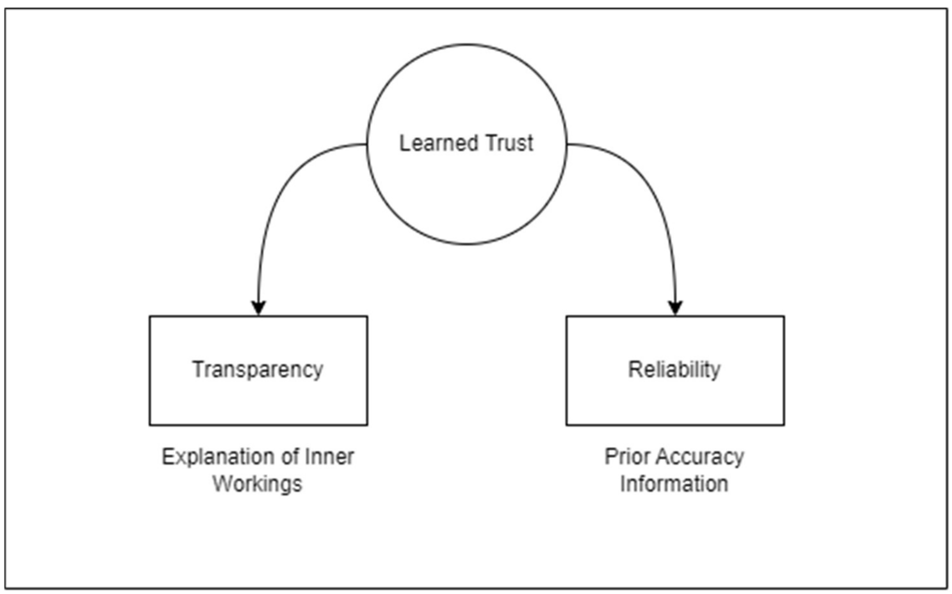

Figure 2. Factors of Trust in Automation
Adapted from Hoff & Bashir (2015)

Transparency involves providing users with greater information regarding how the AI operates. I operationalize the concept of transparency by providing participants with an explanation of the AI model. Demystifying the AI with details regarding the explicit model, factors, and its processes can aid users in understanding the capabilities and context in which the AI can appropriately function. With this increased understanding, auditors theoretically can improve appropriate reliance on AI by knowing when the factors, process, and model match their data and informational needs. Likewise, through interacting with the AI and learning about its reliability, auditors can be exposed to the system's performance capabilities and observe how well it has functioned in the past. Objective details such as the performance history of the AI across several past engagements can help build confidence in auditors about the AI's ability to continue performing the task well.

To summarize, the interaction between trust and algorithm bias involves several factors, some of these being: reliability, perceived competence of the algorithm, and its transparency or understandability. These factors can either positively or negatively affect reliance on algorithms. For instance, if individuals perceive an algorithm as more competent than humans, their trust in the algorithm increases. Conversely, perceiving the algorithm as less competent than humans diminishes trust. Furthermore, misunderstanding an algorithm can lead to divergent outcomes. On one hand, some may infer from the algorithm's complexity that its processes are superior, thereby increasing trust. On the other hand, others might view this complexity as overly risky, reducing their trust. The intent of integrating algorithms extends beyond eliciting appreciation or aversion; the goal is to achieve calibrated, appropriate reliance. This "calibration" ensures that the algorithm is integrated thoughtfully into judgment and decision-making processes. Measuring proper calibration of an AI tool for a subjective and complex estimate is extremely difficult, and as such, this study only investigates the algorithm aversion aspect of algorithm bias. Regardless, the concept of calibration is worth discussing to better understand the context and other concerns involving algorithm bias.

Calibration describes the relationship between an individual's trust in automation and the automation's capabilities (Lee and See, 2004). When calibration is perfect, individuals will trust automation to the exact capabilities of an algorithm. Additionally, an individual's use of a system "will be better… when one's trust is well calibrated to the actual trustworthiness of the system" (Seong & Bisantz, 2008, p. 611). Accountants, especially auditors, are expected to use their training to exercise professional skepticism and evaluate information critically. If accountants display algorithm appreciation, they may not be able to identify mistakes from an AI tool due to over-reliance when not warranted. However, if accountants display algorithm aversion, they may lose potential efficiency and effectiveness gains because of under-reliance on the AI tool. These two types of biases reflect both the risk of not catching errors generated by the AI tool as well as wasting spurning improved audit quality offered by the decision aid. Implementing different interventions to counteract algorithm aversion could help improve accountants' judgment and decision-making processes.

From the Hoff and Bashir (2015) conceptual model, it can be inferred that trust in automation has an inverse relationship with algorithm aversion. The higher an individual's trust in automation, the less the individual will display a negative bias toward automation. At the extreme however, individuals may display too much trust in automation, leading them to rely on it without appropriate professional skepticism. Auditors face potentially high litigation and regulatory risks, translating into severe negative consequences if they blindly rely on audit evidence without evaluating its sufficiency and appropriateness. Lowe et al. (2002) provide evidence that jurors found auditors more responsible for errors when they overrode a highly reliable decision aid. Audit firms would likely only implement AI technology if it was highly reliable, thus exposing auditors to more responsibility and blame if they overrode the system. Too much trust in technology may lead to over-acceptance, while too little trust may lead to under-acceptance. Reliability and transparency of the AI's process help improve an individual's understanding of the algorithm and their confidence in the algorithm's capabilities to perform the assigned task.

2.3 Reliability

Accounting research has investigated how decision-aid reliability influences professionals' judgment and decision-making processes. Providing explicit information about reliability has yielded mixed effects on reliance (Gomaa et al. 2011; Alvarado-Valencia and Barrero 2014; Commerford et al. 2022). However, explicit information about reliability is designed to increase trust between the user and the system (Kaplan et al. 2023). My study further investigates whether providing reliability information increases reliance on AI through a mechanism of increased trust.

Explicit reliability information has produced mixed results on how individuals rely on algorithmic decision aids. Users seem to increase reliance on financial decision aids as the reliability increases (Gomaa et al. 2011). However, once the reliability reaches a certain point, the individuals wholly rely on the decision aid instead of constructively evaluating its advice. This can be problematic, especially as it could lead to failure to detect algorithmic errors (Lee and See 2004). Perceived poor system performance increases aversion, while positive historical performance increases trust in the system (Alvarado-Valencia and Barrero 2014).

AI in auditing should be inherently highly reliable if public accounting firms implement them due to the highly regulated and litigious environment they function in. Despite this, auditors vastly preferred human valuations over AI-generated valuations, even when the AI-generated valuations were more credible and an approved source of audit evidence (Commerford et al. 2022). These mixed findings about how individuals interact with algorithms with explicit reliability underscore the need for additional research on how and when information about explicit reliability could help mitigate algorithm bias in an auditing setting.

Finally, auditors are subject to external pressures such as time and budget pressures, litigation risk, and the requirement to justify decision aid reliance (Ashton 1990; Lowe et al. 2002; Gomaa et al. 2011; Jung and Seiter 2021). Auditors face considerable external pressures during an audit engagement, as managers and partners create time- and financial-based incentives to work quickly to stay under budget. Ashton (1990) demonstrates that justification and financial incentives negatively affect the use of decision aids. However, in the Ashton (1990) study, the aid was only 50 percent reliable, providing users with a tool that was no better than guessing. Time pressure strongly affects the use of decision aids and seems to mitigate algorithm aversion (Gomaa et al. 2011; Jung and Seiter 2021). Additionally, users who viewed the decision to use the aid as more internally motivated than externally motivated displayed higher levels of reliance (Kaplan et al. 2001). The audit field is characterized by many external pressures (budgets, time constraints, working with teams, and short engagement periods), which could lead to an over-reliance on AI even when not appropriate. This over-reliance may cause auditors to not exercise appropriate professional skepticism and fail to detect algorithmic errors.

Historical performance is a prime factor underlying a system's reliability, which refers to "the consistency of an automated system's functions," and is a critical antecedent of trust (Hoff and Bashir 2015, 424). An individual's trust can be strengthened through consistent evidence of positive results, leading to improved reliance on the system. However, prior studies on the role of reliability in increasing reliance on automation have yielded inconclusive results. Even when the perceived performance difference between the system and humans is insufficient, users still prefer the human counterparts despite any advantages automation offers (Bigman and Gray

2018; Longoni et al. 2019). Thus, even when presented with reliability information, individuals may prefer their own perceptions of the automations' credibility and capability. Notably, in the Bigman & Gray (2018) study, even when automation held a clear advantage in the decision-making process of a moral issue, more than a quarter of participants still preferred the human counterpart. Bigman & Gray (2018) further suggest that limiting the automation to an advisory role increased the participants' reliance on the automation. Thus, even when explicit emphasis is placed on automation's superior performance in subjective tasks, individuals may prefer to use human judgment, perceiving an inherent limitation in an algorithm's ability to make nuanced decisions without concrete data.

By examining this dynamic through the lens of learned trust and positioning AI as an advisor, this study aims to delve deeper into how auditors perceive AI as a decision-making aid. Auditors, often facing extreme external pressures of time and budgetary constraints, may find themselves either over-relying on AI or defaulting to their own judgement over a perceived unreliable aid, potentially leading to suboptimal decisions. Such external pressures can cause individuals to prefer their own judgment over a perceived unreliable aid, potentially leading to a suboptimal decision. Conversely, excessive trust in a highly reliable aid could result in overlooking errors a decision aid generates (Ashton 1990; Gomaa et al. 2011; Lee and See 2004). Providing auditors with historical performance information is posited to enhance their trust in the AI system's capabilities, thereby promoting an appropriate level of reliance on the tool. Reliability information aims to balance auditors' reliance on AI, ensuring that auditors' choice to rely on the AI tool aligns more closely with the system's demonstrated reliability, rather than being swayed unduly by external pressures or inherent biases.

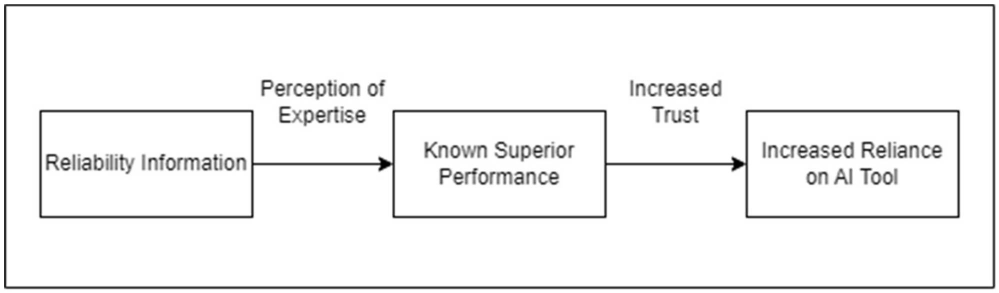

Figure 3. Hypothesized Process between Reliability Mechanism and Reliance

To summarize, the presentation of reliability information is expected to enhance an individual's trust in the AI tool's ability to function effectively, as supported by Dowling & Leech (2007), Solberg et al. (2022), and Kaplan et al. (2023). This trust is anticipated even when participants are not personally familiar with the tool, as the external validation of its positive historical performance can instill confidence in the AI's expertise. If the mechanism functions as anticipated, the expectation of consistent performance should strengthen trust in the AI tool, subsequently leading to greater reliance on it. Figure 3 illustrates the hypothesized process by which historical reliability information influences reliance. Building on this discussion, I propose that auditors who are informed about the AI tool's reliability will be more likely to rely on the system's recommendations than those who have not been provided with such information. Figure 4 graphically depicts the hypothesized relationship The hypothesis is formally stated as follows:

H1: Participants provided with information about the high historical accuracy of the AI tool will exhibit greater reliance on the AI's suggested output relative to participants not provided with the same information.

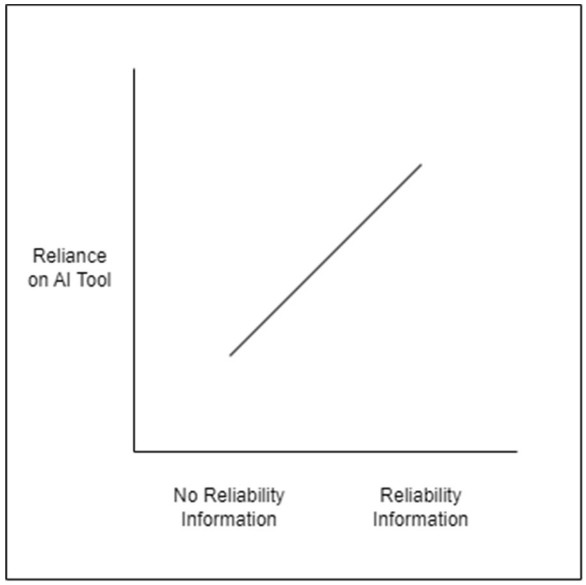

Figure 4. Hypothesized Simple Effect of Reliability Information

2.4 Transparency

I rely on the definition of transparency as “the inner workings or logic [used by] the automated systems are known to human operators to assist their understanding about the system” (Seong & Bisantz, 2008, p. 611). Specifically, I operationalize the concept of transparency through providing additional explainability to the algorithm’s process. Hoff & Bashir (2015) discuss several research studies that provide cognitive feedback and performance feedback about the system’s performance with the label of increased transparency. I believe that one of the fundamental underpinnings of this concept means that transparency (explainability) leads users to have an increased understanding of the system they are interacting with. The increased understanding leads to increased confidence and trust, which in turn increases reliance on the system.

Libby & Luft (1993) provide a framework that explains how knowledge is a key piece of an individual's performance. They state that "much of experienced decision makers' advantage lies in their larger knowledge store and, more importantly, the manner in which they organize their knowledge so that it can be effectively brought to bear on the problem" (Libby & Luft 1993, p. 429). Accounting information systems are often designed to leverage expert knowledge structures by constructing expert systems to assist users. At a meta level, accounting research has investigated how learning about expert systems can potentially improve individual judgment and decision-making processes with the idea that knowing how expert systems function can help users learn better. However, these studies have yielded mixed results. Steinbart and Accola (1994) find that providing explanations for how the process of the expert system did not improve individual performance, increase learning, or increase user satisfaction. Another study suggests that including declarative expert system explanations improved individual learning, while procedural explanations did not affect the learning process (Smedley and Sutton 2004). My transparency intervention seeks to provide participants with an opportunity to learn about the AI's processes and logic, which is a form of a procedural explanation.

Transparency into an algorithm's processes is often obscured—intentionally or accidentally (Langer and Landers 2021). Intentional obscurity may be employed as a management tactic to induce compliance (Schlicker et al. 2021), while unintentional or partial obscurity seems to have an aversion-inducing effect (Shin and Park 2019). Shin and Park (2019) observed that individuals often react negatively when they perceive the inputs or factors of automation as outside their locus of control, leading to a decreased reliance on the system. Furthermore, complex algorithms can diminish reliance, especially when they outperform humans, as individuals may believe the human recommendation process more comprehensible than the algorithm recommendation process (Yeomans et al. 2019). They further speculate that an algorithm's accuracy or reliability is insufficient to increase reliance—the users must be able to understand the process to trust in the system's capabilities.

Based on this understanding, algorithm transparency offers crucial information to users that can foster appropriate trust in algorithms by enhancing understanding of how the system operates, particularly within contexts where reliance is deemed appropriate. Transparency serves as a potential prerequisite to learned trust and affects how the auditor understands the AI's processes. A lack of insight into algorithm's inner workings can lead to "misuse, disuse, and/or abuse" (Hoff & Bashir, 2015, 422). By clarifying or reducing the uncertainty associated with using an AI model, transparency can mitigate these risks (Bhatt et al., 2021). As a form of pre-existing knowledge within the framework of learned trust, an individual's deeper greater understanding of how an algorithm operates is theorized to lead to a greater reliance on the system. Thus, enhancing algorithm transparency not only demystifies the AI's processes but also bridges the gap between user trust and AI capabilities, which potentially increases the effective use of the system.

Detailed explanations of an algorithm provide insight into how it processes information, which could help individuals understand the process and help to calibrate their expectations of how well the algorithm can and should perform (Dzindolet et al. 2003, Seong & Bisantz, 2008). Improving auditors' understanding of the AI's processes should improve reliance on AI systems, contingent on the explanations being sufficient to clarify the system's inner workings. However, if the detailed explanations create confusion, complexity, or cause the problem to seem like it is more difficult, the additional transparency may lead to decreased reliance.

To further delve into algorithm transparency, this study draws on the "Explainable AI" (XAI) literature, which seeks to clarify the rationale behind algorithmic decisions (Van Lent et al., 2004; Biran & Cotton, 2017; Molnar, 2021). Recent auditing literature provides a framework of different types of XAI that can enhance accountants' understanding of algorithms. According to Zhang et al. (2022), XAI techniques primarily fall into two categories: post-hoc and ex-ante. Post-hoc techniques retrospectively analyze an AI's operation, employing complex mathematical models like Partial Dependence Plots or example-based methods such as counterfactual explanations. These methods require large data sets with historical information to validate and verify the AI's processes.

Zhang et al. (2022) utilized these post-hoc methods in their exploration of financial restatements as indicators of audit quality, marking a significant step in employing AI to identify determinants of such quality. This approach mirrors traditional auditing research, which has extensively examined elements like discretionary accruals (Jones, 1991; DeChow et al., 1995) and real earnings management (Roychowdhury, 2006; Kothari et al., 2016) to assess factors leading to financial restatements. These research studies investigated various aspects of financial statements to determine which financial statement items correspond with restatements as a proxy for audit quality. These studies all use prior historical information to back into these factors, including the XAI models. Zhang et al. (2022) provide substantial evidence that AI can perform these same investigations and that various post-hoc techniques can be leveraged to understand how the model arrived at its conclusions.

The complement to post-hoc analysis in Explainable AI (XAI) is ex-ante techniques, which prioritize interpretability by detailing the model's functions and decision-making processes upfront. These techniques range from straightforward models like linear regressions and decision trees, designed to demystify the AI's information processing and output derivation, to more intricate systems like neural networks, aimed at clarifying the determinants of the model’s decisions. In the context of audit procedures that incorporate AI tools during fieldwork, auditing standards mandate the substantiation of these tools both as audit evidence and within audit documentation. This requirement underscores the importance of a clear understanding of the AI models, ensuring that auditors can confidently rely on these tools in compliance with regulatory expectations.

Zhang et al. (2022) further explain the role of feature-driven explanations in adhering to auditing standards, such as AS 1105 and AS 1215. First, they highlight the necessity for auditing firms to precisely identify and document the critical features influencing the AI model's predictions. This clarity not only enhances understanding of the model's prioritized factors but also equips auditors with the justification needed for employing these factors in predictive analyses. Second, explaining the directional impact of these features on outcomes deepens insights into the causality within the AI's processing, shedding light on how input variables influence the final decision. Third, by detailing the operational mechanics of the AI model, such documentation demystifies the "black box" nature of AI, offering a clearer approximation of its functionality. Lastly, understanding the potential modifications in outcomes based on changes in features underscores the dynamic nature of AI decision-making.

In essence, the explicit variable inputs, or features, serve as the foundation for the AI's decisions. Providing auditors with this detailed knowledge fosters the benefit of an enriched understanding of the AI processes. This enhanced comprehension is pivotal in demystifying AI operations, potentially increasing auditors' reliance on the tool, as evidenced in Figure 5.

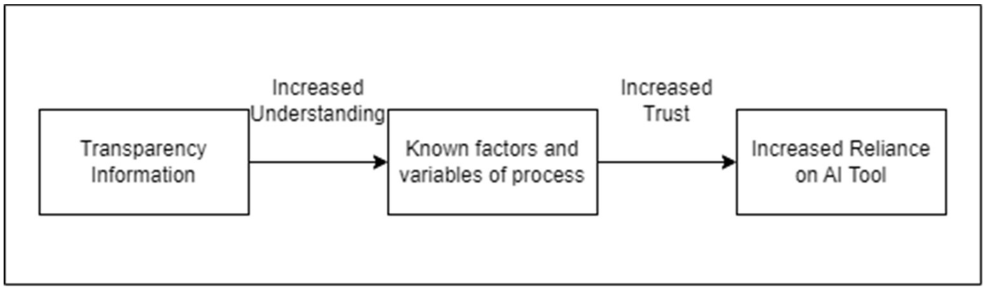

Figure 5. Hypothesized Process between Transparency Mechanism and Reliance

By furnishing participants with specific insights into the variables that influence the AI tool's decisions, this study investigates whether feature-based explanations amplify the understanding of AI's workings. A heightened understanding, in turn, is hypothesized to foster increased reliance on the AI tool relative to conditions offering merely a summary overview of the algorithm's processes. Figure 6 depicts the hypothesized effect and the hypothesis is formally stated as follows:

H2: Participants provided with a feature-driven explanation about the AI's implementation of firm-approved methodologies will exhibit greater reliance on the AI's suggested output relative to participants provided with only a summary explanation.

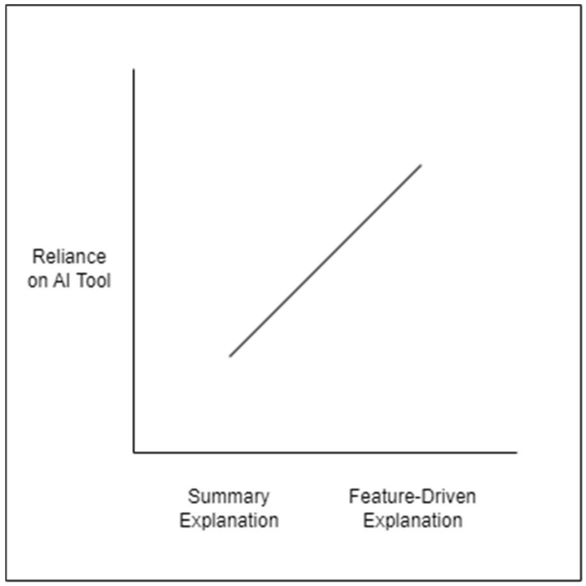

Figure 6. Hypothesized Simple Effect of Transparency Information

2.5 Interactive Effects of Reliability and Transparency on Reliance

While theory supports the predictions of increased reliance on AI through the increased informational value of historical performance and increased understanding and transparency of the AI's processes, it is unknown how these two mechanisms will interact. Hoff & Bashir (2015) explain transparency as increased understanding of a system's reliability but also label the system's reliability as a distinct variable affecting an individual's learned trust. Reliability information has a direct impact on an individual's trust of a system while transparency about a system's processes helps to increase a user's understanding of the system, which in turn should increase reliance. While reliability may have a greater impact on reliance as a first order effect compared to transparency's influence on reliance through understanding as a second order effect, prior research and theory do not predict which intervention would have a stronger impact on users of automation. Therefore, I pose the following research question.

RQ1: Will there be a difference in the extent of reliance on the AI's suggested output between participants receiving only high historical accuracy information and participants receiving only information about the AI's implementation of firm approved methodologies?

Task and individual characters also significantly affect how users approach algorithms. Auditors often face complex tasks requiring appropriate expertise to perform their duties. Public accounting firms address this need by specifically assigning tasks to roles, making sure that complex tasks are performed by staff with sufficient experience. However, task complexity and user expertise seem to be fully mediated by task difficulty (Parkes 2017). When there is an increase in task complexity, users perceive it to be more difficult instead of more complex. However, increasing expertise decreases the perception of the task's difficulty. If AI tools are viewed as experts, then explicit information about reliability should increase the perceived competence of the algorithm. Additionally, auditors seem to strongly prefer a familiar and understandable process, especially for complex tasks (Commerford et al., 2022).

Considering the above literature review and hypotheses development, introducing transparency and reliability should increase reliance. An increase in transparency also increases understanding of the system. Auditor expertise should allow them to believe in the process of AI. Additionally, if the AI process uses similar weights and factors to the human decision-making process, that may increase potential familiarity with the process of the complex judgment. Increasing explicit information about the reliability of the AI should increase trust, thus allowing auditors to have confidence in the competency of the AI. However, these two variables may have a crowding out effect. Figure 7 depicts a graphical representation of RQ1 and RQ2. Positive historical performance and increased understanding of the system could have an additive or substitutive effect on user's reliance of the AI tool. Since there is not a strong theoretical basis to predict an additive effect, I pose the following research question.

RQ2: Will the extent of reliance on AI differ for participants provided with information about both high historical accuracy and information about the AI's implementation of firm-approved methodologies relative to participants provided with information about each type information independently?

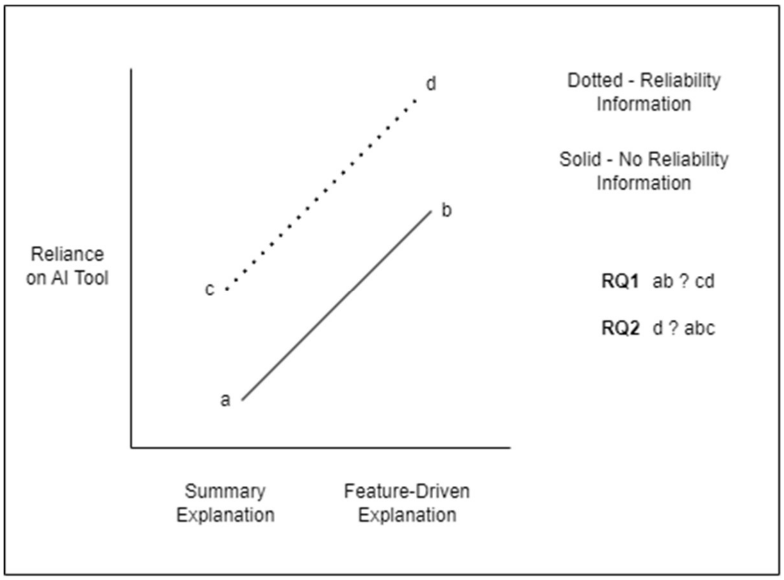

Figure 7. Research Questions and Interactive Effects of Factors of Trust

Chapter Three: Method

3.1 Initial Pilot Study

I ran a pilot study to gather initial data and provide proof of concept for the main study. 36 auditing undergraduate students and two professionals took the initial study. About 45 percent of the participants identified as female and 55 percent were under the age of 26. The participants were presented with background information, an initial demographic question set, the case data adapted from Commerford et al. (2022), and then concluded the study with a post-experimental questionnaire. The pilot study yielded evidence in support of the notion that individuals who receive reliability information about an AI tool rely more on its generated estimate relative to those who did not receive the same information ($p = .027$). However, there was no evidence indicating that transparency influenced the participants to alter their reliance on the AI tool ($p = .467$).

Based on findings from the initial pilot study, three changes were made to the instrument for the main study. First, the overall length of the instrument was shortened as the average time to complete the pilot study was longer than 30 minutes. Recruiting practitioners is expensive and difficult to achieve, especially when professional accountants have busy seasons around the end of the calendar year. By shortening the instrument to an average of 30 minutes, I hoped to increase my likelihood of recruiting practitioners by providing a reasonable compensation for their time of about $50 per hour. Second, the transparency intervention did not seem to be salient to the participants. As a result, a secondary pilot study was launched online via CloudResearch to gauge the effectiveness of different algorithm explanation types. Third, I added a dichotomous variable for the transparency manipulation check. The XAI feature-driven explanation involves explicitly providing factors and variables that the AI tool uses to arrive at its decision. To ensure the saliency of the manipulation, I directly ask the participants if the study provides this information. Originally, the transparency manipulation check only involved measuring pre- and post- understanding of AI models and calculating the difference between the two. However, this measure captures the intended effect of understanding, which may or may not include the participants' direct realization of the provided intervention.

3.2 Secondary Pilot Study

For my subsequent pilot study, I aimed to refine the transparency intervention, given its lack of validation and significant impact on reliance in the initial pilot. To this end, I designed a vignette about a credit card application process facilitated by an artificial neural network (ANN) intending to compare different transparency explanations. This study employed a 4 x 1 design manipulating the transparency explanation across four distinct levels: a strictly academic description of ANN, a process-oriented description of the ANN, an XAI-informed feature-driven explanation, and feature-driven approach with percentage weights to indicate factor importance. The selection of these explanation types was informed by the initial pilot study's outcomes and a review of XAI literature, which suggests varying effects of explanation types on user behavior. The academic and process approach served as the initial interventions, focusing on general model understanding and detailed process insight, respectively. The feature-driven explanations, both with and without weights, were introduced to assess the balance between understandability and detail sufficiency, as these aspects have been highlighted as influential in explanation effectiveness (Hoffman et al., 2018; Zhang et al., 2022).

To evaluate the effects of these four alternative transparency explanations, I recruited 101 participants from the CloudResearch survey website, focusing on their transparency and information content in an AI model explanation. My dependent variable was measured using an 11-point Likert scale, gauging perceptions of transparency, anchored on “Not Very Clear” and “Very Clear”. Additionally, I assessed participants’ perception of how informational the model explanation was. The goal was to distinguish between the clarity of the explanation (i.e., transparency) and the richness or utility of the information provided (i.e., informational). The distinction is crucial to ensure that I captured the participants’ understanding of the model—reflecting transparency, versus their perceptions of the explanation’s utility—indicating information content.

The analysis revealed that the participants preferred the feature-driven explanation significantly to the academic explanation ($p = .050$). While the overall model achieved marginal significance ($p = .086$), the comparison between the other conditions did not display a significant difference in transparency ratings. Interestingly, the ratings for informational content did not vary significantly across any of the conditions, indicating that the study successfully differentiated between the understandability of the explanations and their perceived usefulness.

These findings led to a strategic adaptation of my original transparency intervention of shifting to reflect the XAI feature-driven approach that emphasizes the importance of highlighting which variables and factors the AI tool used in its decision-making process and output. This adaptation aims to enhance the transparency manipulation by providing feature-driven explanations to improve the clarity of how the AI model operates.

3.3 Participants

The main study recruited 39 students from a graduate auditing course at a large public university in the southern United States and 40 professionals from the university's alumni association and advisory board, and the researcher's own professional network. About 31 percent of participants identified as female and 62 percent were under the age of 26. The average external audit experience was 27.5 months, the average internal audit experience was 8.6 months, and the average professional experience was 63.4 months. These descriptive statistics are broken down by condition in Table 1.

Table 1
Demographic Descriptive Statistics

Mean (St. Dev)	**Experimental Condition**				
	Base (n = 12)	**Reliability (n = 18)**	**Transparency (n = 12)**	**Transparency-Reliability (n = 10)**	**Total**
Age (% <Age 26)	58%	63%	67%	80%	62%
Gender (% Female)	25%	19%	25%	60%	31%
External Audit Experience [a]	29.7 (42.7)	23.2 (27.5)	31.5 (69.0)	28.0 (43.8)	27.5 (45.0)
Internal Audit Experience [a]	8.2 (20.5)	11.0 (21.7)	13 (33.1)	0 (0)	8.6 (22.4)
Prof. Accounting Experience [a]	43.3 (58.1)	61.9 (70.4)	82.9 (113.2)	66.8 (86.6)	63.4 (81.6)

[a] Experience measured in months.

* None of these cells are significantly different from one another.

Of the professional participants, 25 of the 40 were manager-level or higher within their firm. Specifically, 11 of these individuals were either directors, partners, or c-suite executives. Professionals who were certified mostly held CPA licenses (19), but some of the individuals also held CISA (2), CFE (1), CMA (1), and EA (1) certifications or licenses.

3.4 Independent Variables

The study involves a 2x2 between-participants design experiment where the manipulated variables are the historical performance of the AI (between-participants; present or not present) and the explanation of the AI's processes (between-participants; summary or feature-driven). I describe the performance of the AI relative to human performance rather than stating the accuracy level of the AI in absolute terms. If one were to state the accuracy level of the AI, the *error rate* would also be implicitly stated (unless the AI accuracy level was set to 100 percent, which is unrealistic). Dietvorst et al. (2015) noted that individuals are particularly sensitive to algorithmic error rates. These individuals who are made aware of algorithmic errors are prone to algorithm aversion. Auditors also seem to judge algorithmic errors more harshly than human errors (Commerford et al., 2022). Therefore, the error rate is not explicitly stated to avoid inadvertently triggering algorithm aversion.

3.5 Dependent Variables

The main dependent variable of interest is the extent to which the participant relies on the AI tool's estimation, or the reliance score. The reliance score is calculated as the audit adjustment proposal divided by the recommended audit adjustment by the AI tool. Scaling the reliance score does not change the variation with the variable, but it does allow for an easier interpretation of the measure. The possible values of the scaled score would range between zero and one.

$$Reliance\ Score = \frac{Participant\ Audit\ Adjustment\ Proposal}{AI\ Tool\ Audit\ Adjustment\ Recommendation} \quad (1)$$

The maximum amount the participants can propose as an audit adjustment is \$20 million. The bank management proposes a \$0 audit adjustment. The reliance score would be calculated as the participant's proposed estimate divided by the total possible adjustment. A \$10 million adjustment would signify that the participant relies equally on management's assertion and the AI tool's assertion, resulting in a 0.5 score for reliance. A \$5 million adjustment would signify that the participant relied more on the management's assertion and would result in a 0.25 score for reliance. A \$15 million adjustment would signify that the participant relied more on the AI tool and would result in a 0.75 score for reliance.

The next dependent variable of interest is a composite reliance measure based on a set of post-experiment questions designed by Parkes (2017). Accounting research has several different methods of measuring reliance on decision aids, including measuring users' agreement with the aid (Hampton 2005), directly asking about their reliance on the aid (Madhavan and Weigmann 2007), and the likelihood of relying on the aid in the future (Hodges et al. 2021; Ganbold et al. 2022).

Rose (2002) notes that measures specifically asking about a user's reliance on an aid potentially fail to grasp important constructs that drive a user's reliance, including prior knowledge, confidence, influence on decision strategy or decision process, user preference (Ye and Johnson 1995), decision aid design (Eining et al. 1997), and potential forced reliance (Kachelmeier and Messier 1990; Glover et al. 1997). Parkes (2017) constructs a combined measure of factors that lead to reliance on technology, including the use of the system in decision-making, the weight given to the DA recommendation, the degree of following recommendations by the DA, integration of DA output (altering decision-making process), use of DA outputs, and if users followed the aid's recommendations (if different than the user's personal opinion). Each factor captures a vital task or individual characteristic that a single measure may fail to capture. These questions were adapted to be specific to the participants' perception of the AI tool's usefulness.

I convert Parkes' reliance constructs into equivalent statements that address various facets of reliance on the AI tool. The first statement attempts to capture how much participants use the AI tool in their decision-making process. The second statement tries to capture how much weight, or importance, the participants assigned to the AI tool. The third statement captures to what degree the participants followed the AI Tool's estimation. The fourth statement asks the participants how much they integrated the output of the AI tool into their judgment and decision-making process. The fifth statement simply asks about how much the participant used the AI tool output. The last statement asks the participants if they followed the AI tool's estimation, *even if* they disagreed with it personally. The scores of each of these statements were averaged together to create the composite reliance measure. The Cronbach's alpha coefficient for the participants' responses to these statements was greater than 0.80, indicating internal consistency, and suggests that the set of questions for the composite score was reliable (Cronbach Alpha = 0.913).

Participants were asked explicitly about their likelihood of relying on the aid for two reasons. First, the question acted as a robustness check, which matches other current accounting research on reliance on AI (Hodges et al. 2021; Commerford et al. 2022; Ganbold et al. 2022). Secondly, it can help to validate Parkes's (2017) designed composite measure for adequate and appropriate use within auditing and accounting research. Validating the Parkes composite measure potentially provides another way of measuring reliance on decision aids and other technological tools within accounting literature.

3.6 Experimental Audit Case

I utilize Commerford et al. (2022) case audit study as the basis for my experiment with four significant changes in the design. First, I only include a firm-provided AI advisor instead of providing a human advisor. Second, instead of varying the objective and subjective inputs into management's decision, I hold the objective information input as a constant, which is explicit to the participant. Their study provides evidence that a difference exists in the perception of objective and subjective inputs, which could influence users' interactions with the AI. The participants relied more on the AI when told that only objective information was used. Third, I introduce a new between-participants intervention of providing explicit performance history of the AI. In the *reliability* condition participants are informed that the AI tool has been successfully used in the past 20 audit engagements and has been 96 percent more accurate than its human counterpart. The *no reliability information* condition only mentions that knowledgeable auditors and third-party experts have internally verified the AI Tool, that the firm considers it to make reasonably accurate predictions, and that it is considered an approved source of audit evidence. Although explicitly stating the new information of performance history would draw specific attention to the manipulation, in a realistic setting, audit firms would want to communicate the potential benefits and strengths of the implemented AI tool to influence higher reliance on the tool. Thus, the manipulation mimics the information auditors receive from their superiors when directed to use an AI tool in judgment and decision-making. Lastly, I introduce a between-participants manipulation of transparency, with the *more transparent* (feature-driven) condition providing information about the relevant and important factors with examples of how the factors impact the output of the AI tool's estimate. The *less transparent* (summary) condition only provides a general description of the AI's model while the XAI feature-driven explanation includes three of the most important variables that the AI tool considers in its calculations (see Appendix A for manipulation conditions). By providing information about the AI's model, the "black box" of the AI's processes should become more transparent and increase the participant's understanding of the AI. The increase in understanding should also result in increased trust in the AI.

3.7 Experimental Procedure

Participants are introduced to the case setting where a fictional audit firm has been engaged to evaluate the loan loss reserve (LLR) estimate as part of the regular audit engagement for a fictional national bank. The participants receive background information about the firm and the bank, including its use of AI in assisting with auditor judgments. The firm states that the AI tool has been internally tested, and they assert a 95 percent satisfactory effectiveness in producing the LLR. I manipulate the presence of the prior performance history of the AI (reliability) in the background information. They assume the role of an in-charge auditor and are responsible for evaluating management's estimates and methodology. The fictional bank utilizes an objective methodology in calculating loan grades for the LLR. The audit firm develops its own loan grades and uses that information to estimate the LLR.

The participants then receive information about the LLR estimate arrived at by the firm and the bank. The LLR information includes a side-by-side report of the firm's and the bank's estimates and methodologies. Additionally, the participants receive information that the bank is confident in its estimates and prefers not to make an adjustment. The LLR estimate arrived at by the AI audit tool would result in an income-reducing audit adjustment of $20 million. The participants decide how much of the adjustment to accept and input their proposed audit adjustment for the LLR. They then decide how much of the proposed adjustment will be accepted by the bank's management after the engagement team's leadership meets with the bank's management. I included this second question to determine if participants believe that there is a difference between what the LLR *should* be and what the LLR would be agreed upon. Participants may have experience with negotiating audit adjustments with clients, which could impact how they propose an estimate. The proposed estimate is supposed to capture what the participants believe is the "true" value of the LLR while providing an opportunity to answer the negotiated estimate, if there happens to be a difference.

The reliability intervention includes further information that the AI tool has been integrated into the past 20 audit engagements and has been 96 percent more accurate than their human counterparts' estimations. The detailed feature-driven transparency condition includes a description of variables the model considered important for its estimation process with specific examples of how variation in these features would affect its output (see Appendix A). The summary transparency condition only mentions that the AI Tool system uses an artificial neural network without describing its processes or models.

The proposed adjustment measures serve as a dependent variable of the extent to which participants relied on the AI, indicating how much of the proposed audit adjustment they recommend ($0 indicates fully accepting management's estimate, $20 million indicates fully accepting the AI's estimate). I also include a measure that asks participants to provide their expected audit adjustment value, if the participants anticipate a difference between the two values. However, as Rose (2002) points out the fallibility of using participant agreement alone as a measure of reliance, I include two alternate dependent variables in the post-experiment questionnaire measured on a nine-point Likert Scale. The first is a composite of indirect measures derived from Parkes (2017) that asks about different constructs that lead to reliance on technology. The second alternative dependent variable is a direct measure that asks the participant to indicate how likely they are to rely on an AI tool in the future to make these decisions, consistent with the approach employed in recent accounting research measuring reliance (Hodges et al. 2021; Commerford et al. 2022; Ganbold et al. 2022).

Other questions captured general demographic information, work experience, familiarity with technology, perspectives of AI use within their professional work, and the freedom to exercise their judgment in using AI, if applicable. The mediators capture the participants' perception of evidence quality from either the AI or management, their perception of objectivity of management and AI, their general confidence in AI, and their confidence in AI to perform the LLR calculation. These mediators are designed to capture potential processes that may provide insight and clarification into the relationship between the experimental interventions and the various measures of participant reliance on the AI tool.

Chapter Four: Results

4.1 Manipulation and Attention Checks

To assess the effectiveness of the experimental interventions, I conducted specific checks of the reliability and transparency manipulations. For the reliability manipulation check, a majority (67%) of the participants correctly assessed the presence of the intervention, indicating a successful manipulation. Only 46% of the participants correctly answered the transparency manipulation check, which inquired if there was a description of the AI tool's factors. This suggests that the participants may have conflated other narrative elements with the intended transparency intervention. Despite the high failure rate for the transparency manipulation check, there was a significant increase in self-reported understanding ($p = .009$) in the transparency condition, suggesting that the intervention increased participants' understanding of AI models.

The full sample of 79 participants failed to generate any significant results in the statistical model ($p = .201$). After identifying the professional/student status as a covariate, a marginally significant model emerged for assessing the participants' reliance score ($p = .064$). Within the full-sample ANCOVA model, only the reliability condition ($p = .057$) and the professional/student status covariate ($p = .076$) approached significance. The transparency condition ($p = .462$) and the combined transparency-reliability condition ($p = .363$) were insignificant. However, after removing the 27 individuals who had failed the attention check (described next), I was able to obtain a more significant statistical model ($p = .009$).

The attention check required the participants to correctly identify the AI tool's purpose. Twenty-seven of the participants incorrectly answered the attention check. Notably, about half of the student sample (18 out of 39) failed the attention check. Additionally, 16 of the 27 individuals failed the transparency manipulation check and 14 of these individuals failed the reliability manipulation check. Only one individual failed all three checks.

Excluding the participants who failed the attention check reduced the sample to 52 participants (31 professionals and 21 students) for hypothesis testing who were distributed across the four conditions: base (12), reliability (18), transparency (12), and combined transparency-reliability (10). Refining the sample for the hypothesis testing helped to increase the validity of the findings and interpretations of this study.

4.2 Analysis Methodology

I test my hypotheses with a 2x2 analysis of covariance (ANCOVA) model with auditors' proposed adjustments divided by total possible adjustment as the main dependent variable (reliance score), reliability and transparency as the factors, and the participant's status as a professional or student as a covariate. Two alternate dependent variables are the composite reliance measure and the likelihood of reliance, which are based on self-reported measures in the post-experimental questionnaire. Results reported throughout this section are for the main dependent variable, reliance score, unless otherwise noted. For Research Questions 1, I use a planned contrast test to analyze the potential difference in the effect of either the reliability intervention or transparency intervention of the AI tool. For Research Question 2, I use a planned contrast test to analyze the potential interaction effect between reliability and transparency on participant reliance of the AI tool.

Participants' pre-understanding of AI as a covariate does not seem to influence the results of the study ($p = .789$) nor does the length of their professional experience ($p = .512$). However, there is a significant difference in responses to the main dependent variable when the participants were divided by professional/student status ($p = .041$). As suggested by Piercey (2022), I also investigate the interactions of the professional/student status (PS) with the independent variables. None of the various interactions between the independent variables (reliability and transparency) and the PS variable are significant ($p = .183$), and when including these interactions, the PS covariate approaches significance ($p = .053$). I include PS as a covariate for my reliance score analyses which utilize an ANCOVA model. The alternative dependent variables, namely the composite reliance measure and the likelihood of reliance measure, are analyzed with ANOVA models since none of the potential covariates are significantly correlated with these dependent variables.

I use Levene's Test of Equality and the F Test for Heteroskedasticity to test the ANCOVA and ANOVA assumptions of normally distributed residuals. While Levene's Test is significant ($p = .025$), the F Test is not significant ($p = .276$). Because the Levene's Test is significant, I ran the Kolmogrov-Smirnov and Shapiro-Wilk normality tests. The former is not significant at $p = .161$ but the Shapiro-Wilk test is significant at $p = .002$. The alternative dependent variables' results are not normally distributed per the normality tests as well. Given the small numbers in my study ($n = 52$), with an unequal distribution of participants in each condition, the ANCOVA and ANOVA models do not seem to be robust to violations of the normality assumptions. As such, I perform nonparametric tests to substantiate my findings of the dependent variables within the ANCOVA and ANOVA models.

Nonparametric models are not bound by the normality assumptions required for ANCOVA and ANOVA models, offering a flexible and robust alternative for data analysis. A key distinction is their reliance on group medians instead of means for statistical testing. This makes the nonparametric models less susceptible to outliers and skewed distributions, providing a more accurate estimate of the data's central tendencies. The Kruskal-Wallis test, a nonparametric approach, is particularly suited for situations where there are three or more groups to compare, each group is independently sampled, and the data does not meet parametric test prerequisites. Given that the data from this study meets these criteria, I have opted for the Kruskal-Wallis test as an alternative to validate the findings from my ANCOVA and ANOVA models under conditions of violated normality assumptions. While nonparametric tests may not be as powerful as their parametric counterparts, such as ANOVA models, they are useful for confirming trends when parametric assumptions are not met.

The nonparametric analysis of each dependent variable yielded significant p-score results that align closely with those from the ANCOVA and ANOVA models in my tests (reliance score, $p = .036$; composite reliance measure, $p = .041$; likelihood of reliance, $p =.729$). Consequently, the nonparametric findings are consistent with ANCOVA and ANOVA models, and despite the violations of normality. I report the ANCOVA and ANOVA models' findings to remain consistent with prior accounting literature.

4.3 Tests of Hypotheses and Research Questions

Table 2 presents the reliance score variable, descriptive statistics, ANCOVA tests, and contrast tests. The reliance score is scaled, which is calculated by dividing the participants' proposed audit adjustment by the potential total adjustment ($20 million), and ranges from 0 (no reliance on the AI tool estimate) to 1 (complete reliance). The base condition showed the lowest mean reliance score ($\bar{x}$ = .20). In contrast, the reliability ($\bar{x}$ = .54), transparency ($\bar{x}$ = .47), and combined transparency-reliability ($\bar{x}$ = .57) conditions displayed notably higher mean scores. Specifically, conditions featuring reliability information ($\bar{x}$ = .55) yielded higher mean reliance scores compared to those without ($\bar{x}$ = .33). Within the transparency interventions, the feature-driven approach ($\bar{x}$ = .51) outperformed the summary approach ($\bar{x}$ = .40), leading to an overall mean reliance score of $\bar{x}$ = .45 across all conditions. These findings indicate a tendency for participants to place slightly greater trust in the AI tool's LLR estimate over the management's estimate, highlighting the influence of reliability intervention to increase reliance on the AI tool.

Table 2
Hypotheses Tests: Reliance Score on AI Tool Estimate

Panel A: Main Reliance Score - Mean (Standard deviation) [N] Cell

		Transparency - Summary	Transparency – Feature	Overall Row
Reliability of AI Tool	None	0.20 (0.17) [12] A	0.47 (0.35) [12] B	0.33 (0.30) [24]
	Present	0.54 (0.37) [18] C	0.57 (0.31) [10] D	0.55 (0.345) [28]
	Overall Column	0.40 (0.35) [30]	0.51 (0.33) [22]	0.45 (0.34) [52]

Table 2 (continued)

Panel B:
ANCOVA Table

Source of variation	df	Mean Square	F	*p*-value
Reliability Information (H1)	1	0.534	5.666	0.021*
Transparency Information (H2)	1	0.152	1.610	0.467
Transparency-Reliability	1	0.141	1.497	0.227
Professional/Student	1	0.417	4.422	0.041*
Error	47			

Panel C: Contrast Test

	df	t	*p*-value
B = C (RQ1)	1, 37.87	2.449	0.018*
D > ABC (RQ2)	1, 47	1.303	0.198

Notes:
*Significant at the p = .05 level. Two-tailed p-values are reported.
**Levene's Test significant at p = 0.025; Nonparametric significant at p = .036
***Contrast tests do not assume equal variance.

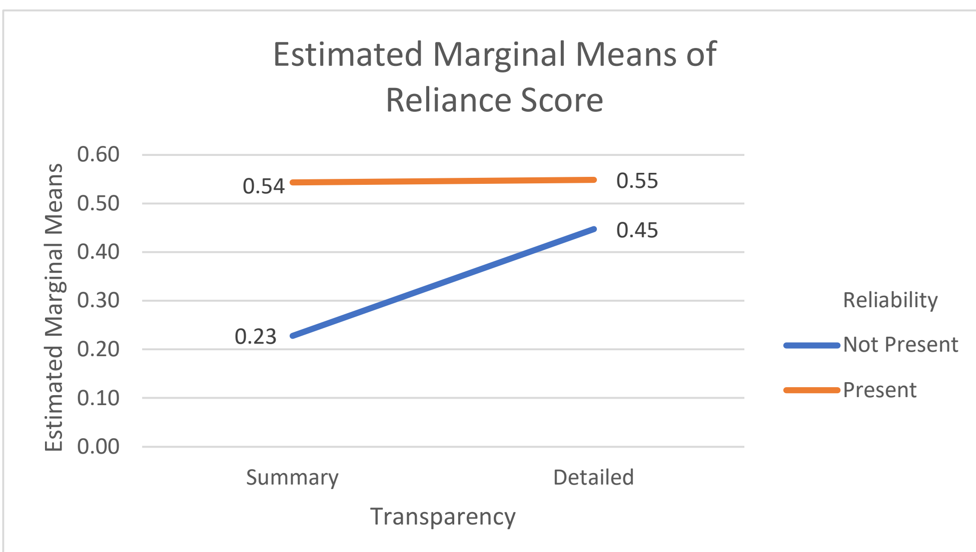

Figure 8. Effects of Interventions on Auditor Proposed Estimate
Reliance score standardized by dividing participant estimate by total possible estimate
See Table 2 for reliance score information and Appendix C for non-scaled information.

I further investigate the difference in behavior between the professionals and students with respect to the reliance score dependent variable. The professionals' results were similar to the complete sample's results. Specifically, the reliability intervention significantly increased reliance on the AI tool and was significantly different than the transparency intervention (See Table 3).

While these are certainly effects of a low power statistical analysis and small sample size (Professionals – N =31, Students – N = 21), the pattern of responses between the two groups are visually distinct. The professionals' responses suggest that additional reliability information and transparency information seem to increase reliance on the AI tool (See Figure 9).

Table 3
Hypotheses Tests: Reliance Score on AI Tool Estimate (Professionals)

Panel A: Main Reliance Score - Mean (Standard deviation) [N] Cell

		Transparency - Summary	Transparency - Feature-Driven	Overall Row
Reliability of AI Tool	None	0.16 (0.17) [9] A	0.24 (0.34) [6] B	0.19 (0.24) [15]
	Present	0.49 (0.40) [11] C	0.58 (0.43) [5] D	0.51 (0.40) [16]
	Overall Column	0.34 (0.35) [20]	0.40 (0.40) [22]	0.36 (0.37) [31]

Table 3 (continued)

Panel B: ANOVA Table

Source of variation	df	Mean Square	F	*p*-value
Reliability Information (H1)	1	0.774	6.612	0.016*
Transparency Information (H2)	1	0.054	0.461	0.503
Transparency-Reliability	1	0.000	0.003	0.955
Error	27			

Panel C: Contrast Test

	df	t	*p*-value
B = C (RQ1)	1, 27	2.571	0.016*
D > ABC (RQ2)	1, 27	1.413	0.102

Notes:
*Significant at the p = .05 level. Two-tailed p-values are reported.

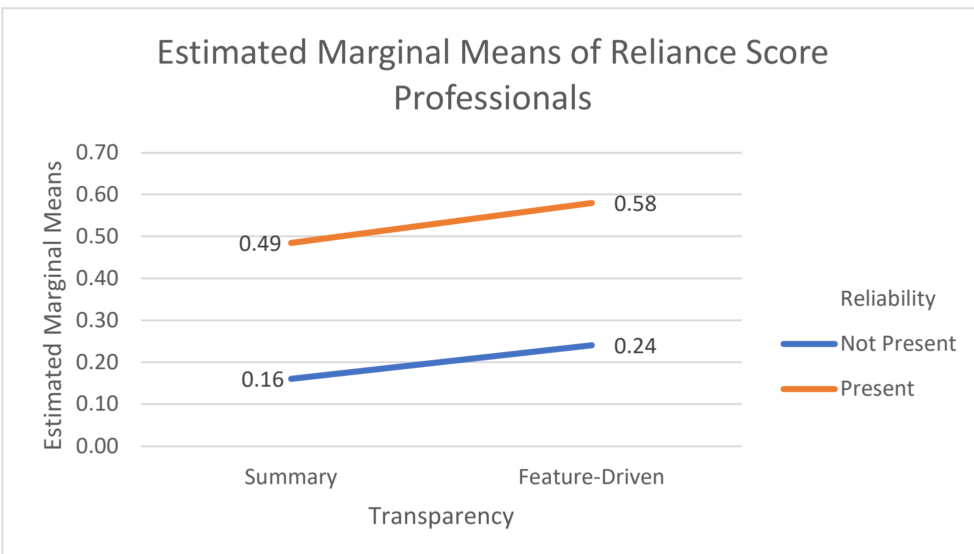

Figure 9. Effects of Interventions on Auditor Proposed Estimate - Professionals
Model approaches significance at p = .086.
See Table 3 for further information.

The students' behavior, when modeled in ANOVA, is statistically insignificant and prevents any specific conclusions from being drawn from the subsample (see Table 4). However, the students' responses suggest that transparency information alone seems to increase reliance while combining both reliability and transparency information seems to decrease reliance (See Figure 10).

Table 4
Hypotheses Tests: Reliance Score on AI Tool Estimate (Students)

Panel A: Main Reliance Score - Mean (Standard deviation) [N] Cell

		Transparency - Summary	Transparency - Feature-Driven	Overall Row
Reliability of AI Tool	None	0.32 (0.13) [3] A	0.69 (0.17) [6] B	0.57 (0.24) [9]
	Present	0.63 (0.33) [7] C	0.55 (0.17) [5] D	0.60 (0.27) [12]
	Overall Column	0.53 (0.31) [10]	0.63 (0.18) [11]	0.58 (0.25) [21]

Panel B: ANOVA Table

Source of variation	df	Mean Square	F	*p*-value
Reliability Information (H1)	1	0.036	0.647	0.432
Transparency Information (H2)	1	0.106	1.926	0.183
Transparency-Reliability	1	0.238	4.308	0.053
Error	27			

Panel C: Contrast Test

	df	t	*p*-value
B = C (RQ1)	1, 17	2.571	0.432
D > ABC (RQ2)	1, 17	0.069	0.946

Notes:
*The model is insignificant at p = .183

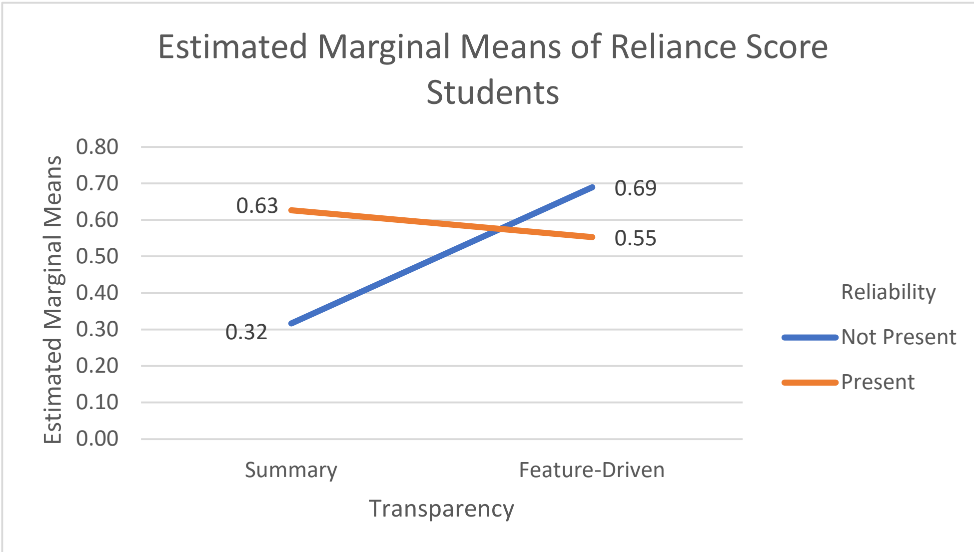

Figure 10. Effects of Interventions on Auditor Proposed Estimate - Students
Model is not significant at p = .183
See Table 4 for further information.

These results help provide an explanation for why the status of the participant (professional or student) acted as a covariate in the main model on a limited basis. The professionals' behavior seems to trend in the hypothesized direction of results, which contributes towards the complete sample's results. Namely, their reliance increased when provided with reliability information and visually increased (while remaining statistically insignificant) when provided with transparency information. However, the students' behavior only partially trended towards the hypothesized direction of results. While the visual pattern of means indicates that their reliance increased when provided with reliability information and transparency information independently, providing both sources of information simultaneously seems to have a conflicting effect on reliance. Due to low statistical power given the sample size, this interpretation should be taken lightly until further studies can provide a more robust analysis.

Descriptive statistics, ANOVA tests, and contrast tests for the composite reliance measure variable are detailed in Table 5. The base condition shows the lowest mean score ($\bar{x}$ = 4.29), indicating minimal reliance. In comparison, the reliability condition demonstrates a significantly higher mean ($\bar{x}$ = 6.14), followed by the combined transparency-reliability condition ($\bar{x}$ = 5.82), and the transparency condition alone ($\bar{x}$ = 5.09). Conditions featuring reliability information have a higher overall mean ($\bar{x}$ = 6.02) compared to those without it ($\bar{x}$ = 4.60). The summary transparency condition ($\bar{x}$ = 5.40) slightly outperforms the feature-driven transparency approach ($\bar{x}$ = 5.33), with an aggregate mean of $\bar{x}$ = 5.37 across all conditions. These results highlight the significant impact of reliability information on enhancing reliance, with transparency interventions showing no significant effect on the composite reliance measure.

Table 5
Hypotheses Tests: Composite Parkes Measure on AI Tool Estimate

Panel A: Likelihood of Reliance - Mean (Standard deviation) [N] Cell

		Transparency - Summary	Transparency - Feature	Overall Row
Reliability of AI Tool	None	4.29 (1.18) [12] A	4.92 (2.30) [12] B	4.60 (2.05) [24]
	Present	6.14 (1.87) [18] C	5.82 (1.56) [10] D	6.02 (1.75) [28]
	Overall Column	5.40 (2.04) [30]	5.33 (2.01) [22]	5.37 (2.01) [52]

Table 5 (continued)

Panel B: ANOVA Table				
Source of variation	df	Mean Square	F	*p*-value
Reliability Information (H1)	1	23.422	6.391	0.015
Transparency Information (H2)	1	0.285	0.078	0.782
Transparency-Reliability	1	2.785	0.760	0.388
Error	48			

Panel C: Contrast Test			
	df	t	*p*-value
B = C (RQ1)	1, 48	2.528	0.015*
D > ABC (RQ2)	1, 48	1.037	0.305

Notes:

*Significant at the p = .05 level

**Model is marginally significant at p = 0.060

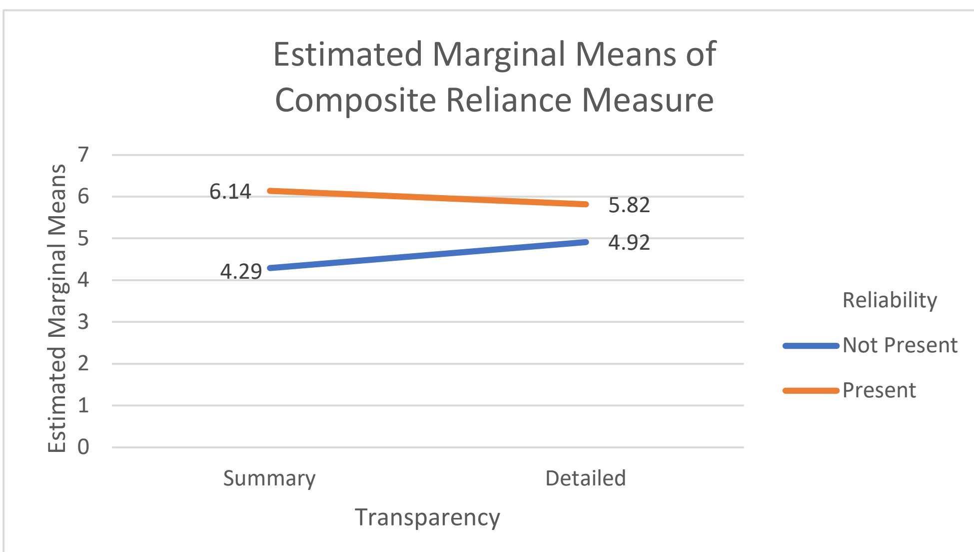

Figure 11. Effects of Reliability and Transparency on Composite Reliance Measure

Composite reliance measure derived from Parkes (2017) measures

See Tables 3 and 6 for further information.

Descriptive statistics, ANOVA tests, and contrast tests for the Likelihood of Reliance variable are shown in Table 6. Participants assessed their future likelihood of utilizing an AI tool for LLR estimates on a 9-point Likert scale, ranging from strongly disagree to strongly agree. Across the experimental conditions the likelihood of reliance measure did not reveal any statistically significant differences, even with the inclusion of covariates such as experience, professional/student status, or initial understanding, all of which fail to enhance the model's significance ($p = .652$).

The base condition reports the lowest mean likelihood of reliance ($\bar{x} = 4.83$). In comparison, the reliability condition shows a higher mean ($\bar{x} = 5.67$), and the combined transparency-reliability condition has the highest ($\bar{x} = 5.90$), with the transparency condition alone closely mirroring the base ($\bar{x} = 4.82$). Conditions incorporating reliability information exhibit a greater mean likelihood of reliance ($\bar{x} = 5.75$) compared to those that do not ($\bar{x} = 5.21$). Among the transparency intervention, the feature-driven approach results in a slightly higher mean ($\bar{x} = 5.73$) compared to the summary approach ($\bar{x} = 5.33$), leading to an overall mean of $\bar{x} = 5.50$ across conditions. This suggests that, on average, participants hold a fairly neutral stance towards the future use of an AI tool for making LLR estimate decisions.

Table 6
Hypotheses Tests: Likelihood of Reliance on AI Tool Estimate

Panel A: Composite Parkes Measure - Mean (Standard deviation) [N] Cell

		Transparency - Summary	Transparency - Feature	Overall Row
Reliability of AI Tool	None	4.83 (2.21) [12] A	5.58 (2.02) [12] B	5.21 (2.02) [24]
	Present	5.67 (2.50) [18] C	5.90 (1.29) [10] D	5.75 (2.12) [28]
	Overall Column	5.33 (2.38) [30]	5.73 (1.70) [22]	5.50 (2.11) [52]

Panel B:
ANOVA Table

Source of variation	df	Mean Square	F	*p*-value
Reliability Information (H1)	1	4.104	0.898	0.348
Transparency Information (H2)	1	3.001	0.656	0.422
Transparency-Reliability (H4)	1	0.828	0.181	0.672
Error	48			

Panel C: Contrast Test

	df	t	*p*-value
B = C (RQ1)	1, 48	0.947	0.348
D > ABC (RQ2)	1, 48	0.714	0.479

Notes:
*Model is not significant

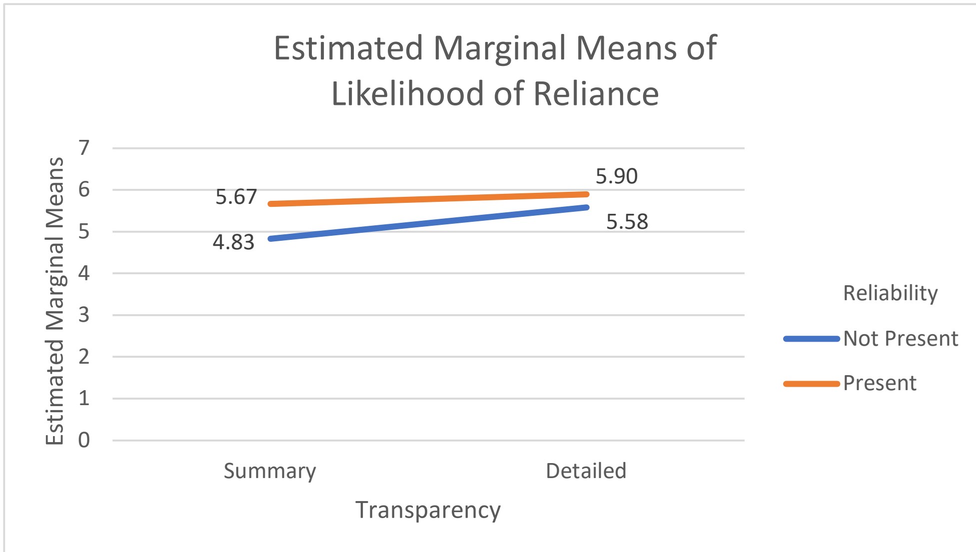

Figure 12. Effects of Reliability and Transparency on Likelihood of Reliance
Likelihood of reliance measures willingness to use AI tools in future decisions
See Table 6 for further information.

Results of the study provide evidence to support Hypothesis 1 (H1). H1 predicts a main effect from providing reliability information to participants, which would result in higher reliance on the AI tool than the base condition. The reliance score variable (Table 2, Panel B) is significant at p = .021 ($F_{1,\ 47}$ = 5.66), suggesting that providing reliability information increases reliance on the AI tool. The composite reliance measure (p = .015; Table 5, Panel B) also supports this finding while the self-reported likelihood of reliance (p = .348; Table 6, Panel B) is insignificant and does not support H1. Thus, the overall results provide evidence suggesting that providing reliability information increases participants' reliance on the AI tool.

Next, I discuss my tests of hypothesis for Hypothesis 2 (H2). Overall, the discussion of these results should take into consideration the low statistical power of the analysis, strongly encouraging a cautious interpretation of the results. H2 predicts a main effect from providing transparency information to participants, resulting in higher reliance on the AI tool than the base condition. The reliance score ($p = .467$; Table 2, Panel B), composite reliance measure ($p = .782$; Table 5, Panel B), and likelihood of reliance score ($p = .422$; Table 6, Panel B) are insignificant. There is no main effect from providing transparency information about the AI's process, suggesting that participants do not increase their reliance on the AI tool when provided with detailed feature-driven information about how the AI functioned. The results do not support H2.

Research Question 1 (RQ1) investigates the potential difference between the reliance displayed by participants in the reliability and transparency conditions. I analyzed this research question using a contrast test (reliability = 1, transparency = -1). While I expected the reliability and transparency conditions to improve participant reliance, the results seem to suggest there is a significant difference between the two interventions. The reliance score was significant ($p = .018$, two-tailed; $t = 2.449$; Table 2, Panel C), the composite reliance measure was significant ($p = .015$, two-tailed; $t = 2.528$; Table 5, Panel C), and the likelihood of reliance was not significant ($p = .348$, two tailed; Table 6, Panel C). These results suggest that the reliability condition influenced participants to rely on the AI tool to a greater extent than the transparency condition.

Research Question 2 investigates the potential difference in reliance between transparency-reliability condition and the other conditions. I used a planned contrast test where I coded the base, reliability, and transparency conditions as -1 and the combined transparency-reliability condition as 3. None of the dependent variables are significant, with the lowest p-value greater than $p = .198$ (reliance score). Each contrast test can be found in Panel C of their respective tables. These results suggest that the combined condition did not improve reliance on the AI tool more than the other conditions combined. Thus, I am unable to answer RQ2.

4.4 Additional Analyses

4.4.1 Parkes' Composite Measure Individual Constructs

Table 7 shows the descriptive statistics for each question of the composite reliance measure that I adapted from Parkes (2017). There is no statistically significant association between any of these constructs and the reliance score dependent variable, only at the composite-level which is discussed above with respect to H1 and RQ1.

Table 7
Descriptive Statistics of Post-Experiment Questionnaire Composite Reliance Measure[a,b]

Mean (St. Dev)	**Condition** Base (n = 12)	Reliability (n = 18)	Transparency (n = 12)	Transparency-Reliability (n = 10)
Use of System in Decision Making	5.2 (2.0)	6.6 (2.1)	6.3 (2.1)	6.7 (1.7)
Weight Given to System	4.7 (2.4)	6.7 (2.6)	5.3 (2.8)	6.5 (1.9)
Degree of Following System	3.8 (2.3)	6.2 (2.5)	4.3 (2.9)	5.8 (2.4)
Integration of System Output (Altered JDM)	4.5 (2.1)	6.2 (2.4)	4.8 (3.1)	5.0 (2.2)
Use of System Outputs	4.9 (2.2)	7.1 (2.0)	5.8 (2.7)	6.5 (1.8)
Follow System, even if disagreed	2.8 (1.5)	4.1 (2.6)	2.9 (2.3)	4.4 (2.2)

[a] The post-experiment questionnaire measures on based on an 9-point Likert Scale.
(1 - Strongly Disagree, 5 - Neither Agree nor Disagree, 9 - Strongly Agree)

[b] Adapted from Parkes (2017)

4.4.2 Source Credibility Theory

Source Credibility Theory (Birnbaum & Stegner, 1979) suggests that individuals interpret the credibility of evidence through three factors: source bias, judge bias, and expertise. I capture participants' perceptions of these factors with the following measures: evidence quality of the AI tool or management, bias of AI tool or management, and confidence in AI tools generally or to be able to specifically perform a LLR calculation. These measures are a proxy for the factors proposed by Birnbaum and Stegner (1979) which I further test through a series of process model analyses. In prior accounting research, Cohen et al. (2022) find that investors' perceived effectiveness (credibility) of auditor committees is strongly affected by source competence and source bias. Thus, the participants in this study potentially assessed the effectiveness of the AI tool on similar factors regarding perceived competence (expertise), evidence quality (source bias), and objectivity (judge bias).

As part of the post-experimental questionnaire, I had participants answer three sets of questions to capture these potential mediating and moderating variables. The first set of questions dealt with the participants' perception of the evidence quality presented by the AI and by the bank's management. If participants viewed one set of evidence as having higher quality than the other, that could influence their subsequent decision to rely on the AI tool. The next set of questions asked participants about the perceived level of objectivity presented by the AI and by management. If the participants believed that one party was highly biased, that could also strongly influence why they chose to (not) rely on the AI tool. The last set of questions asked participants to indicate their confidence in the AI to perform well in a general sense and in the ability of AI to generate an accurate LLR estimate. Participants can be confident of AI generally;

however, they may not be confident in AI to perform a specific task. I present the mean and standard deviation of these measures in Table 8.

Overall, participants seemed to perceive the AI tool as having a higher level of objectivity ($\bar{x}$ = 7.1) than management ($\bar{x}$ = 4.5). Participants given both reliability and transparency information seemed to perceive management as less objective ($\bar{x}$ = 4.9) than just the reliability condition alone ($\bar{x}$ = 3.8) and the transparency condition alone ($\bar{x}$ = 4.7). The participants in the base condition perceived management ($\bar{x}$ = 3.8) as much less objective than the AI tool ($\bar{x}$ = 6.8) and providing reliability or transparency information seemed to increase their perception of objectivity of both management and the AI tool.

Table 8
Descriptive Statistics of Post-Experiment Questionnaire[a]

Mean (St. Dev)	**Condition** Base (n = 12)	Reliability (n = 18)	Transparency (n = 12)	Transparency-Reliability (n = 10)
Evidence Quality - AI	5.2 (2.0)	5.9 (2.1)	5.5 (1.9)	6.6 (1.0)
Evidence Quality - Management	5.7 (1.9)	5.2 (1.7)	5.3 (2.1)	6.0 (1.5)
Objectivity - AI	6.8 (2.2)	7.2 (2.7)	7.0 (2.4)	7.5 (1.2)
Objectivity - Management	3.8 (1.9)	4.7 (1.9)	4.7 (1.5)	4.9 (1.4)
Confidence in AI (General)	5.4 (2.3)	5.9 (2.3)	6.0 (2.1)	6.1 (1.2)
Confidence in AI (LLR)	5.1 (2.1)	6.4 (1.8)	6.0 (2.1)	6.0 (1.2)

[a] The post-experiment questionnaire measures on based on an 9-point Likert Scale.

Analysis of the source objectivity variable across conditions revealed no significant variation in participants' perception of bias by either management or by the AI tool. An independent samples t-test comparing participants' perception of objectivity between management and the AI tool demonstrated a significant perception of higher objectivity for the AI tool than management (Table 9). The t-tests confirm significant differences in perceptions of source bias, with participants viewing the AI as being significantly more objective than management ($p < .05$).

Interestingly, the perceived objectivity difference did not translate into significant differences in reliance across the experimental conditions ($p > .10$). This finding suggests that while the participants found the tool to be more objective, that did not materially affect their reliance on the AI tool. Additionally, there was no significant difference between the perceived quality of evidence provided by the AI tool or management across conditions, with the means and t-test results detailed in Table 7.

Table 9

Mean Difference between AI and Management Bias and Evidence Quality by Conditions[a]

	Condition			
Mean Difference [T-Statistic]	**Base (n = 12)**	**Reliability (n = 18)**	**Transparency (n = 12)**	**Transparency-Reliability (n = 10)**
Bias	2.9 [3.495]	2.4 [3.169]	2.3 [2.847]	2.6 [4.549]
Evidence Quality	-0.5 [0.626]	0.7 [-1.125]	0.3 [-0.307]	0.6 [-1.068]

[a] The post-experiment questionnaire measures on based on an 9-point Likert Scale.

* A positive score indicates that the AI score mean was higher than the Management score mean.

4.4.3 Mediation and Moderation Process Models

Next, I discuss the results of mediation and moderation analyses aimed at exploring whether the relation between transparency and reliance is mediated through intervening variables or moderated by other variables. I utilize the Process Macro for SPSS, as recommended by Hayes (2018), to facilitate these analyses.

The initial analysis begins with a serial mediation model (Process Model 6), which assesses the impact of perceived general confidence in AI and specific confidence in AI's ability to perform LLR estimates on the reliance score. This approach is justified by the premise that confidence in AI for the LLR estimates (a specific task) falls within the broader context of general confidence in AI. According to Figure 13, the transparency intervention does not significantly impact general confidence in AI ($\beta = 0.312$, $p = .59$). However, a significant positive relationship exists between general AI confidence and confidence in AI's LLR estimate capabilities ($\beta = 0.77$, $p < .001$), which, in turn, positively influences the composite reliance measure ($\beta = 0.76$, $p = .001$). These findings indicate that although the transparency intervention does not directly affect general or specific task-related confidence in AI, the level of general confidence in AI significantly shapes the confidence in AI's performance of specific tasks like LLR estimates. Importantly, it is this specific confidence in AI's task performance, rather than a broad confidence in AI, that ultimately impacts reliance on the AI tool.

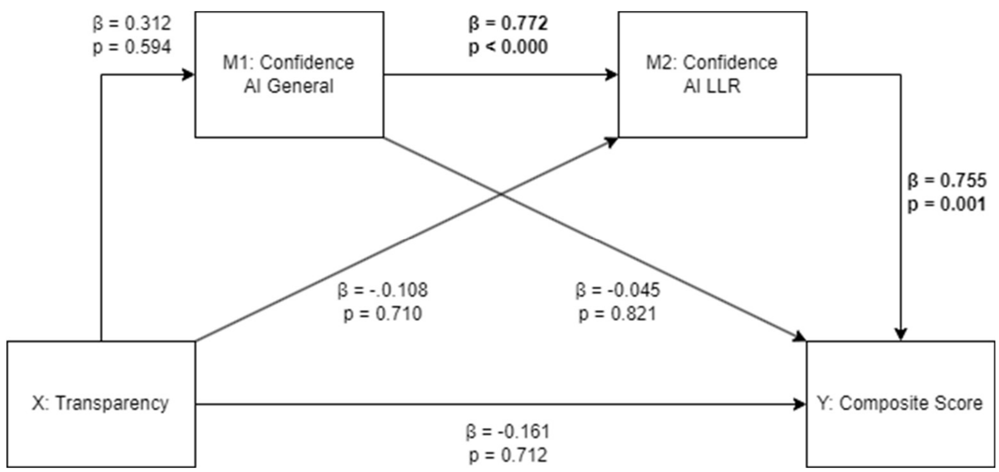

Figure 13. Serial Mediation Analysis on Composite Reliance Measure

This diagram represents a serial mediation analysis (Hayes, 2018). I use a Process Model 6 with two mediators. Specifically, the effect of transparency may operate sequentially through general confidence of AI and confidence in the AI's ability to make an accurate estimate.

In the next analysis, I conducted a parallel mediation using Process Model 4 to explore how perceived evidence quality affects participants' reliance scores (refer to Figure 14). Given that the evaluations of the AI's and management's evidence quality function independently, these variables were analyzed in parallel, contrasting with the serial approach of the previous analysis. The analysis revealed that the transparency intervention did not significantly affect perceptions of the AI's evidence quality ($\beta = 0.34$, $p = .530$). Conversely, a positive relationship was observed between the AI's evidence quality and the reliance score ($\beta = 0.07$, $p = .003$), while participants' evaluation of management's evidence quality was negatively associated with reliance ($\beta = -0.05$, $p = .045$). These findings indicate that participants' reliance on the AI tool's output is influenced by their assessment of the evidence quality provided by both the AI and management. Specifically, higher perceived quality of AI evidence correlates with increased reliance on the AI tool, whereas greater perceived quality of management evidence is correlated with decreased reliance.

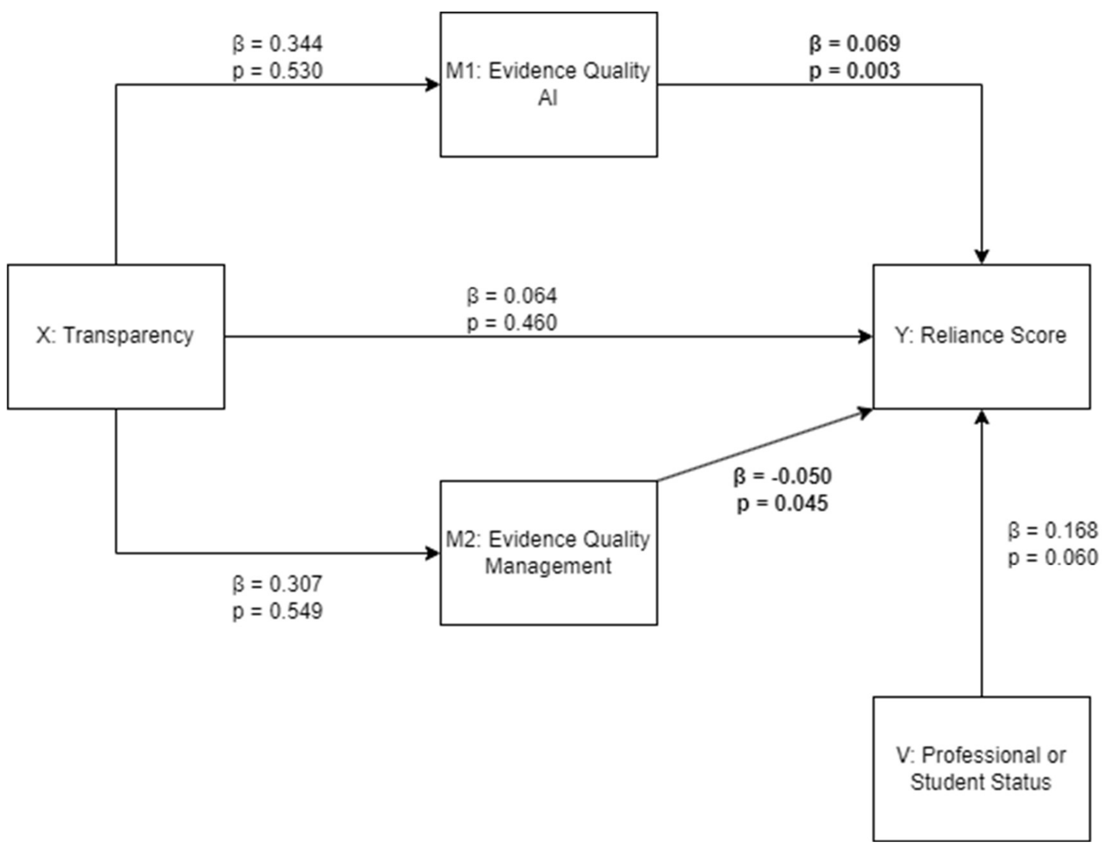

Figure 14. Parallel Mediation Analysis on Reliance Score

This diagram represents a parallel mediation analysis (Hayes, 2018). I use a Model 4 with two mediators. Specifically, the effect of transparency may operate in a parallel fashion through perceived Evidence Quality of the AI and perceived Evidence Quality of the WRB Management.

In a further analysis, I explored the influence of transparency on reliance scores through a moderated mediation framework, employing Process Model 89 with two serial mediators, as illustrated in Figure 15. The analysis centered on the reliance score as the dependent variable, with the transparency intervention as the independent variable, and general confidence in AI alongside confidence in AI's capability for LLR estimates as mediators. The reliability intervention served as the moderator. A key finding from this analysis is that the reliability intervention significantly diminished the effect of confidence in AI's LLR estimate capabilities on the reliance score, rendering it insignificant. This finding highlights the moderating role of reliability, suggesting that it can attenuate the impact of individual confidence in the AI tool's LLR estimation abilities, particularly within the context of the transparency intervention's process model.

The participants were asked to self-report their understanding of AI, artificial neural networks, and machine learning before and after the main portion of the experiment to determine if there is any learning that occurs within the study. Through a simple mediation model (Figure 16), the transparency intervention is positively and significantly associated with an increase in understanding ($\beta = 1.764$, $p = .006$). However, the increased understanding is only marginally significant with an increase in the reliance score ($\beta = 0.037$, $p = .090$) while the transparency intervention is not significant ($\beta = 0.008$, $p = .938$). Only the covariate of PS was significant ($\beta = .188$, $p = .048$), suggesting that the professional/student status of the participants directly influenced reliance score in this specific process model.

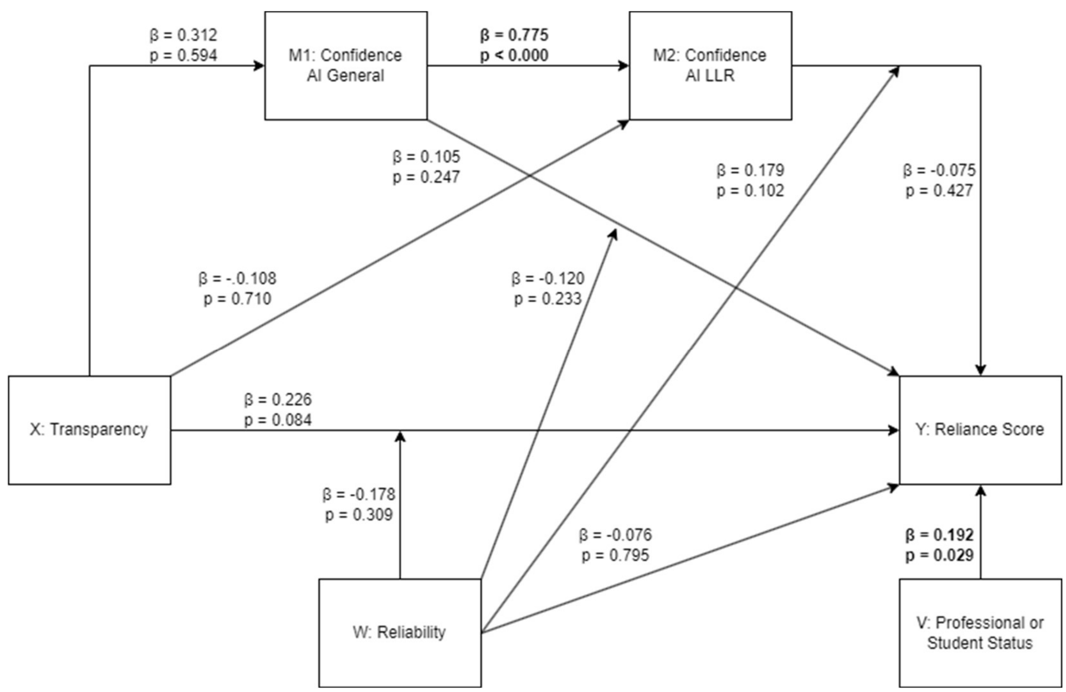

Figure 15. Moderated Mediation Analysis on Reliance Score

This diagram represents a moderated mediation analysis (Hayes, 2018). I use a Model 89 with two mediators. Specifically, the effect of transparency may operate sequentially through general confidence of AI and confidence in the AI's ability to make an accurate estimate. Additionally, the effect of reliability may strengthen or weaken the effects of transparency or the effects of the mediators.

The transparency condition significantly improved participants' self-reported understanding of AI models, as indicated by the notable difference between pre- and post-tests ($p < .01$; Figure 16). However, the increase in understanding did not correspond to increased reliance on the AI tool. In an exploratory analysis, adjusting the process model's confidence interval to 90 percent altered the confidence interval range for the difference in understanding to (.000, .138) from the previous range of (-.008, .151) at a 95% confidence interval. This adjustment suggests that with a larger sample size, the improved understanding attributable to the transparency intervention could potentially serve as a significant mediator in explaining the discrepancy between expected and observed effects of the transparency intervention on reliance.

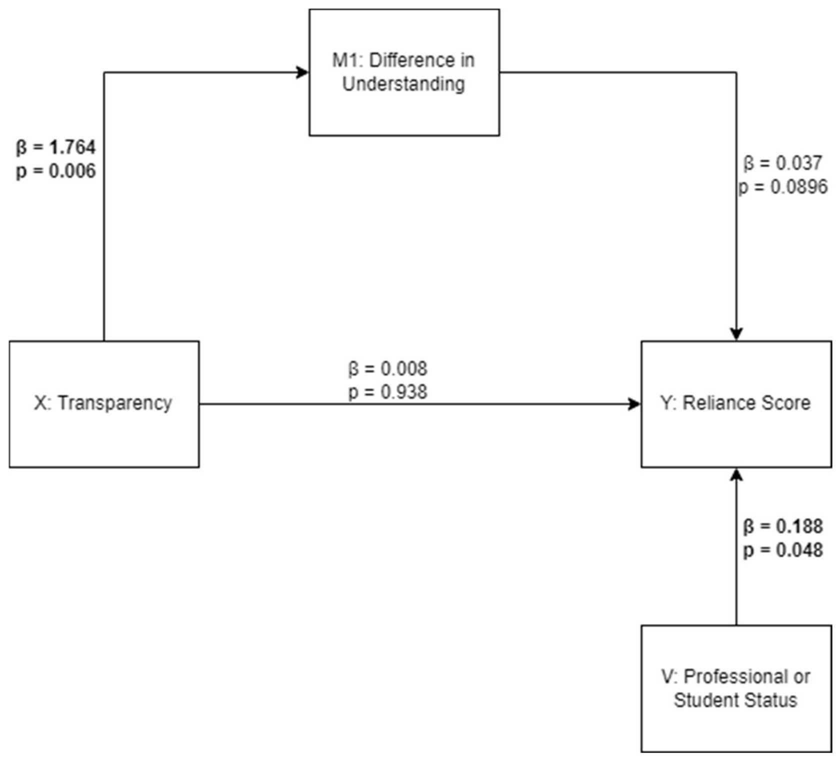

Figure 16. Mediation Analysis of Difference in Understanding on Reliance Score

This diagram represents a parallel mediation analysis (Hayes, 2018). I use a Model 4 with one mediator. Specifically, the effect of transparency seems to increase participants' understanding. However, this understanding does not translate into an increase in reliance on the AI tool.

The intent of the transparency condition was to inform the participants about how the AI tool functioned and increase their understanding of the tool. An increased understanding would help the participants feel comfortable with the tool's processes and increase reliance. Prior studies provide evidence that accountants perform better in structured environments, which is somewhat similar to the case study presented to them in this experiment (Glover et al. 1997). Participants were provided with step-by-step instructions and information about the task and process of the AI tool. Additionally, decision aids with expert-like structures learn more and improve their performance (Rose et al. 2012). However, even though the study provided information about the AI tool's expert validation as well as variables considered important in this complex estimate, the additional detail was insufficient to increase the participants' reliance on the tool. Perhaps the participants may have preferred to use other input components than the ones specified, indicating that they may not have trusted the assumptions the tool may have made.

Decision support system research supports this suggestion with evidence that auditors with prompts and frameworks perform better than those being presented with all the information simultaneously (Murthy & Wheeler 2018). Other research suggests accountants struggle to overcome prior learned biases even when presented with newer, updated information (Beck et al. 2014). It is likely that the participants did not have much experience with an artificial neural network and were unsure of how to incorporate the additional information into their existing knowledge structures. Thus, despite the increased self-reported understanding of AI models, the participants did not correspondingly increase reliance on the AI tool's output.

Chapter Five: Discussion and Conclusion

AI is an especially salient topic with the recent advent of ChatGPT and other advanced AI tools that accounting firms have begun to implement into their normal processes. It is important to understand how to facilitate the appropriate use of AI in a rapidly evolving technological business environment. Accountants face increasing amounts of data and information that can be difficult to use or interpret without suitable tools. However, AI tools offer accountants a new way of improving their ability to interpret large amounts of data and information effectively and efficiently. Reliability and transparency are two potential interventions that firms can implement to improve accountants' reliance on AI tools.

Consistent with my first hypothesis, providing historical reliability information about the positive performance of an AI tool increased participant reliance on the tool. This reliance was measured in two ways – first, by their decision to accept more of the tool's LLR estimate than the participants in the other experimental conditions; second, by their self-reported perceptions of their reliance on the tool, usefulness of the tool, and alignment of opinion on the tool's output. This is informative because the findings suggest that reliability information not only impacts their perceptions of AI but also their behavior.

Inconsistent with the second hypothesis, including transparency information (i.e. important factors and variables that the AI tool incorporates into its decision) does not seem to influence any aspect of the judgment and decision-making process for auditors and their reliance on the AI tool. Additionally, there does not seem to be an additive or multiplicative effect from providing auditors with both reliability and transparency information. This should be interpreted with caution due to the low statistical power in the analysis. Further research should still investigate the mechanism of transparency (i.e. additional explainability) and its influence on auditor reliance on algorithms with a more robust sample size to provide definitive results.

This study makes several contributions to theory and literature. First, this study investigates two potentially significant mechanisms of trust and its impact on algorithm aversion through the lens of learned trust (Hoff and Bashir, 2015). It identifies reliability information as a critical factor that signals an algorithm's credibility and competence, thereby encouraging auditors to place greater reliance on AI tools. This finding enriches the discourse on algorithm aversion, pinpointing a concrete mechanism that fosters reliance on algorithms and addressing a gap in the auditing literature concerning algorithm aversion. My research provides novel insights into the underlying mechanisms influencing individuals' use of algorithms, suggesting that while reliability information may increase reliance, transparency information does not seem to directly impact reliance on algorithms. However, transparency does increase the understanding of AI models, which has a marginal impact on reliance. This distinction provides a unique insight into other mechanisms that might drive algorithm aversion. Furthermore, this research contributes to the information systems literature by providing evidence about how learned trust mechanisms influence algorithm aversion and suggests that reliability information is a key factor for enhancing user interface design and information presentation to promote algorithm reliance.

From a practical standpoint, this study underscores the significance of conveying AI reliability information to auditors, demonstrating that knowledge of positive historical outcomes can bolster tool reliance. Incorporating reliability information emerges as a cost-effective strategy for mitigating algorithm aversion, offering a straightforward method for accounting firms to enhance auditors' trust in AI tools. Conversely, the study suggests that transparency—

though not directly boosting reliance—remains crucial for understanding. However, increasing auditors' understanding of AI processes may not lead to improved reliance on AI tools.

Lastly, the study sheds light on the differing perceptions and uses of AI tools among various user groups within the accounting profession, from interns and staff to decision-makers like directors and partners. It establishes a foundational understanding that professionals and students may rely on AI tools differently, providing valuable insights for firms considering AI implementation and for regulators contemplating requirements for AI tool comprehension among accountants.

The results of this study should be interpreted in light of its limitations, which provide opportunities for future research. Firstly, the sample size of 52 participants is relatively small, with uneven distribution across conditions. Although statistical adjustments were made to mitigate this issue, a study with a larger and more evenly distributed sample—particularly with more professionals—would allow for a more rigorous analysis. Recruiting 40 practitioners during the period of December to January limited participation, which is a time marked by holidays, vacations, and peak financial reporting audit tasks. Given that the working status (professional or student) emerged as a significant covariate, increasing the sample size in both categories could reveal deeper insights into the differential interactions of professionals and students with the AI tool. Since students often serve as proxies for staff accountants, understanding these dynamics could enrich both the academic literature and practical applications, offering insights into how staff, seniors, and potentially directors or partners perceive and utilize AI tools in audit practices.

An additional limitation of this study is that students failed the manipulation and attention checks at a much higher rate than the professionals. Possible reasons for the higher failure rate could be due to differences in incentives between the professionals and students. Professionals were compensated with a $25 gift card for their participation while the students only received extra credit for their coursework. Future research can investigate the dynamic of working status of participants and their reaction towards a specific type of incentive. For example, when individuals are presented with the same task, does extra credit (or other types of course credit) incentivize accounting students to produce the same levels of effort and attention from accounting professionals that are compensated monetarily for their time? If it is an issue of incentives, that could be informative to the academy in prescribing the use of students as proxies for staff accountants and clarify the context for when accounting students behave like professionals, with respect to incentives in completion of research studies.

While this study sheds light on algorithm aversion and the quest to bolster reliance on AI tools that enhance decision-making quality, it only addresses part of the issue related to algorithm bias. Future research can investigate mechanisms that mitigate excessively high reliance on algorithms (algorithm appreciation) and strategies to overcome auditors' hesitancy to rely on beneficial decision aids, addressing the issue of algorithm aversion. A promising avenue for exploration is the concept of algorithm calibration which aims to fine-tune users' reliance on algorithms. Effective calibration would allow auditors to discern when algorithms are the most suitable decision-making tool and when human judgment should prevail. Given the prevalence of subjective decisions and estimates in auditing and accounting, such as the loan loss reserve estimate explored in this study, achieving optimal reliance is complex. Neither algorithms nor human decision-makers are infallible. Future research can focus on investigating mechanisms that can modulate reliance, reducing it among those overly dependent on algorithms while fostering greater trust and utilization among those skeptical of algorithmic outputs.

While this study's exploration of a transparency intervention did not significantly enhance auditor reliance on the AI tool, the reasons for this outcome present avenues for future research. One potential factor is auditors' reservations about the AI's methodology, including concerns over data usage or the relevance of the features the tool utilized for its estimates. For example, one participant noted that they reduced reliance on the tool due to its reference to property value changes in a high-cost area like Seattle, despite the bank's location in the Midwest, where the average property value of buildings and other loans potentially is significantly different. This discrepancy suggests that auditors might question the tool's applicability to their specific context, leading to reduced trust.

Moreover, while an increase in understanding was reported, the extent of this enhancement may not have been substantial enough to alter reliance on the AI tool. Another participant reported their audit firm often overuses "buzz words and fluff" that seemed disingenuous and obfuscating to the actual AI tools' processes and outputs. This led to the individual's belief that "it is easier for me to trust [someone] who I know is biased," which then allowed the individual to take the perceived bias of management into consideration rather than trust the AI tool in this study.

Additionally, divergences in proprietary methodologies across firms, from the Big 4 to national and regional firms, might contribute to skepticism regarding the AI tool's processes. These firms may employ distinct approaches for calculating loan loss reserves, potentially causing perceived misalignment with the AI tool's identified key features.

Addressing these issues calls for future research into diverse XAI methodologies, exploring both ex-ante techniques like decision trees and post-hoc methods such as counterfactual explanations. Tailoring XAI techniques to better match auditing practices could foster a more harmonious integration of AI tools, enhancing their perceived reliability. Additionally, refining how AI processes are explained—moving away from ambiguous terminology towards clearer, more concise explanations—may mitigate concerns about AI being a "black box." Investigating various XAI interventions could also illuminate how different approaches impact auditor understanding and reliance, identifying the most effective strategies for bridging the gap between AI tools and their users.

Another limitation concerns the instrument's length and level of detail. Although designed to capture critical audit information and decision-making processes for external validity, the instrument's portrayal of AI tool processes may not align with the diverse methodologies employed by different public accounting firms, particularly in complex estimates like the LLR calculation. Professionals' familiarity with their firm's specific methodologies may lead to discrepancies in how the AI tool's processes are perceived, while students might lack the necessary background to critically assess detailed client and audit information. Future research could benefit from exploring variations in professional opinions across different firms—categorized by size, region, or other relevant criteria—to identify significant methodological differences. Such findings could pave the way for developing a generalized framework to understand how auditing practice approaches complex estimates.

A natural extension and future research project would be to employ a design science approach or field study to analyze actual AI tools within an accounting context. Partnering with firms would allow me to design potential interventions within an actual tool. Behavioral research requires assumptions and simplifications, which limits potential direct real-world applications. A design science approach would allow me to incorporate the generalizations and new understanding derived from this study into actual work processes. Such an approach would help validate the results of this study and provide further insight and knowledge about the process of how accountants approach AI tools, as well as how to help accountants incorporate them better into their professional duties.

Another extension and future research project would be to analyze other variables within the learned trust layer of Hoff and Bashir's model. They identify more than a dozen different variables that affect how individuals rely on automation, which adds complexity and requires extensive research in determining how these variables interact with one another, and ultimately, how individuals rely on automation and AI. Especially since transparency, which is one of the identified constructs within the learned trust layer, did not directly influence auditor behavior, future studies can investigate the interactive effect of learning and other mechanisms that may a strengthen or weaken the influence of transparency on reliance.

References

AICPA-CIMA. "Artificial Intelligence is a Game Changer for Auditors." AICPA-CIMA. Accessed April 10, 2023. https://www.aicpa-cima.com/news/article/artificial-intelligence-is-a-game-changer-for-auditors.

Alvarado-Valencia, Jorge A., and Lope H. Barrero. "Reliance, trust and heuristics in judgmental forecasting." *Computers in Human Behavior* 36 (2014): 102-113.

Arnold, Vicky, Philip A. Collier, Stewart A. Leech, and Steve G. Sutton. "Impact of intelligent decision aids on expert and novice decision-makers' judgments." *Accounting & Finance* 44, no. 1 (2004): 1-26.

Ashton, Robert H. "Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification." *Journal of Accounting Research* 28 (1990): 148–80. https://doi.org/10.2307/2491253.

Bhatt, Umang, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, et al. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–13. AIES '21. New York, NY, USA: *Association for Computing Machinery*, 2021. https://doi.org/10.1145/3461702.3462571.

Bigman, Yochanan E., and Kurt Gray. "People are averse to machines making moral decisions." Cognition 181 (2018): 21-34.

Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." In *IJCAI-17 workshop on explainable AI (XAI)* Vol. 8, No. 1 (2017): 8-13).

Birnbaum, Michael H. and Steven E Stegner. "Source credibility in social judgment: Bias, expertise, and the judge's point of view." *Journal of Personality and Social Psychology*, 37.1 (1979): 48.

Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann. "Task-dependent algorithm aversion." *Journal of Marketing Research* 56, no. 5 (2019): 809-825.

ChatGPT, response to "Revise this paragraph for language and readability," OpenAI, March, 15, 2024, chat.openai.com/chat.

Commerford, Benjamin P., Sean A. Dennis, Jennifer R. Joe, and Jenny W. Ulla. "Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence." *Journal of Accounting Research* 60, no. 1 (2022): 171–201. https://doi.org/10.1111/1475-679X.12407.

Cohen, J.R., et al. "The effects of audit committee ties and industry expertise on investor judgments—Extending Source Credibility Theory." *Accounting, Organizations and Society*, 102 (2022): 101352

de Vries, Peter, and Cees Midden. "Effect of indirect information on system trust and control allocation." *Behaviour & Information Technology* 27, no. 1 (2008): 17-29.

Dechow, Patricia M., Richard G. Sloan, and Amy P. Sweeney. "Detecting earnings management." *The Accounting Review* (1995): 193-225.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology*: General 144, no. 1 (2015): 114.

Dowling, Carlin, and Stewart Leech. “Audit Support Systems and Decision Aids: Current Practice and Opportunities for Future Research.” *International Journal of Accounting Information Systems* 8, no. 2 (June 1, 2007): 92–116. https://doi.org/10.1016/j.accinf.2007.04.001.

Dzindolet, Mary T., Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. “The Role of Trust in Automation Reliance.” *International Journal of Human-Computer Studies*, *Trust and Technology*, 58, no. 6 (June 1, 2003): 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.

Eining, Martha M., Donald R. Jones, and James K. Loebbecke. "Reliance on decision aids: An examination of auditors’ assessment of management fraud." *Auditing: A Journal of Practice & Theory* 16, no. 2 (1997).

Ganbold, Odkhishig, Anna M. Rose, Jacob M. Rose, and Kristian Rotaru. “Increasing Reliance on Financial Advice with Avatars: The Effects of Competence and Complexity on Algorithm Aversion.” *Journal of Information Systems* 36, no. 1 (Spring 2022): 7–17. https://doi.org/10.2308/ISYS-2021-002.

Glikson, Ella, and Anita Williams Woolley. “Human Trust in Artificial Intelligence: Review of Empirical Research.” *Academy of Management Annals* 14, no. 2 (July 2020): 627–60. https://doi.org/10.5465/annals.2018.0057.

Glover, Steven M., Douglas F. Prawitt, and Brian C. Spilker. "The influence of decision aids on user behavior: Implications for knowledge acquisition and inappropriate reliance." *Organizational Behavior and Human Decision Processes* 72, no. 2 (1997): 232-255.

Gomaa, Mohamed I., James E. Hunton, Eddy H. J. Vaassen, and Martin A. Carree. “Decision Aid Reliance: Modeling the Effects of Decision Aid Reliability and Pressures to Perform on Reliance Behavior.” *International Journal of Accounting Information Systems* 12, no. 3 (September 1, 2011): 206–24. https://doi.org/10.1016/j.accinf.2011.02.001.

Hampton, Clark. "Determinants of reliance: An empirical test of the theory of technology dominance." *International Journal of Accounting Information Systems* 6, no. 4 (2005): 217-240.

Hayes, Andrew F. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression based Approach.* New York: Guilford Publications, 2018.

Hodge, Frank D., Kim I. Mendoza, and Roshan K. Sinha. “The Effect of Humanizing Robo-Advisors on Investor Judgments*.” *Contemporary Accounting Research* 38, no. 1 (2021): 770–92. https://doi.org/10.1111/1911-3846.12641.

Hoff, Kevin, and Masooda Bashir. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human Factors* 57, no. 3 (May 2015): 407–34. https://doi.org/10.1177/0018720814547570.

Hoffman, Robert R., et al. “Metrics for explainable AI: Challenges and prospects.” *arXiv preprint arXiv:1812.04608* (2018).

Hollander, Myles, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. John Wiley & Sons. 2013.

Jones, Jennifer J. "Earnings management during import relief investigations." *Journal of Accounting Research* 29.2 (1991): 193-228.

Jung, Markus, and Mischa Seiter. “Towards a Better Understanding on Mitigating Algorithm Aversion in Forecasting: An Experimental Study.” *Journal of Management Control* 32, no. 4 (December 1, 2021): 495–516. https://doi.org/10.1007/s00187-021-00326-3.

Kachelmeier, Steven J., and William F. Messier Jr. "An investigation of the influence of a nonstatistical decision aid on auditor sample size decisions." *The Accounting Review* (1990): 209-226.

Kaplan, Steven E, J. Hal Reneau, and Stacey Whitecotton. "The Effects of Predictive Ability Information, Locus of Control, and Decision Maker Involvement on Decision Aid Reliance." *Journal of Behavioral Decision Making* 14, no. 1 (2001): 35–50. https://doi.org/10.1002/1099-0771(200101)14:1<35::AID-BDM364>3.0.CO;2-D.

Kaplan, Alexandra D., et al. "Trust in Artificial Intelligence: Meta-Analytic Findings." *Human Factors*, 65.2: 337–359. https://doi.org/10.1177/00187208211013988.

Kapoor, Michael. "Big Four Invest Billions in Tech, Reshaping Their Identities." *Bloomberg Tax*, January 2, 2020. https://news.bloombergtax.com/financial-accounting/big-four-invest-billions-in-tech-reshaping-their-identities.

Kothari, Stephen P., Natalie Mizik, and Sugata Roychowdhury. "Managing for the moment: The role of earnings management via real activities versus accruals in SEO valuation." *The Accounting Review* 91.2 (2016): 559-586.

Koustanaï, Arnaud, Viola Cavallo, Patricia Delhomme, and Arnaud Mas. "Simulator training with a forward collision warning system: Effects on driver-system interactions and driver trust." *Human Factors* 54, no. 5 (2012): 709-721.

Langer, Markus, Cornelius J. König, and Andromachi Fitili. "Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection." *Computers in Human Behavior* 81 (2018): 19-30.

Langer, Markus, and Richard N. Landers. "The Future of Artificial Intelligence at Work: A Review on Effects of Decision Automation and Augmentation on Workers Targeted by Algorithms and Third-Party Observers." *Computers in Human Behavior* 123 (October 1, 2021): 106878. https://doi.org/10.1016/j.chb.2021.106878.

Lee, John D, and Katrina A See. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors*, 2004.

Libby, Robert, and Joan Luft. "Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment." *Accounting, Organizations and Society* 18, no. 5 (1993): 425-450.

Liel, Yotam, and Lior Zalmanson. *What If an AI Told You That 2 + 2 Is 5? Conformity to Algorithmic Recommendations*, 2020.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. "Algorithm appreciation: People prefer algorithmic to human judgment." *Organizational Behavior and Human Decision Processes* 151 (2019): 90-103.

Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge. "Resistance to medical artificial intelligence." *Journal of Consumer Research* 46, no. 4 (2019): 629-650.

Lowe, D. Jordan, and Philip M. J. Reckers. "The Influence of Outcome Effects, Decision Aid Usage, and Intolerance of Ambiguity on Evaluations of Professional Audit Judgement." *International Journal of Auditing* 1, no. 1 (1997): 43–58. https://doi.org/10.1111/1099-1123.00012.

Lowe, D. Jordan, Philip M. J. Reckers, and Stacey M. Whitecotton. "The Effects of Decision-Aid Use and Reliability on Jurors' Evaluations of Auditor Liability." *The Accounting Review* 77, no. 1 (January 1, 2002): 185–202. https://doi.org/10.2308/accr.2002.77.1.185.

Madhavan, Poornima, and Douglas Wiegmann. “Effects of Information Source, Pedigree, and Reliability on Operator Interaction With Decision Support Systems.” Accessed April 15, 2023. https://doi.org/10.1518/001872007X230154.

Makarius, Erin E., Debmalya Mukherjee, Joseph D. Fox, and Alexa K. Fox. “Rising with the Machines: A Sociotechnical Framework for Bringing Artificial Intelligence into the Organization.” *Journal of Business Research* 120 (November 1, 2020): 262–73. https://doi.org/10.1016/j.jbusres.2020.07.045.

Molnar, Cristoph. (2021) Interpretable machine learning. Lulu.com Available at: https://christophm.github.io/intepretable-ml-book/.

Murthy, Uday S., and Patrick R. Wheeler. "The effects of decision-aid design on auditor performance in internal control evaluation tasks." *Journal of Information Systems* 32, no. 2 (2018): 95-113.

Newman, David T., Nathanael J. Fast, and Derek J. Harmon. "When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions." *Organizational Behavior and Human Decision Processes* 160 (2020): 149-167.

Parasuraman, Raja, and Victor Riley. "Humans and automation: Use, misuse, disuse, abuse." *Human Factors* 39, no. 2 (1997): 230-253.

Parkes, Alison. “The Effect of Individual and Task Characteristics on Decision Aid Reliance.” *Behaviour & Information Technology* 36, no. 2 (February 1, 2017): 165–77. https://doi.org/10.1080/0144929X.2016.1209242.

Piercey, M. David. “Throw it in as a Covariate?” Common Problems Using Measured Control Variables in Experimental Research. *Auditing: A Journal of Practice & Theory*, 42.2 (2023): 183-205.

Public Company Accounting Oversight Board (PCAOB). *PCAOB Release No. 2022-002. Planning and Supervision of Audits Involving Other Auditors and Dividing Responsibility for the Audit with Another Accounting Firm*. Washingon, DC: PCAOB, 2022.

Raschke, Robyn L., Aaron Saiewitz, Pushkin Kachroo, and Jacob B. Lennard. “AI-Enhanced Audit Inquiry: A Research Note.” *Journal of Emerging Technologies in Accounting* 15, no. 2 (September 1, 2018): 111–16. https://doi.org/10.2308/jeta-52310.

Riedl, René. "Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions." *Electronic Markets* 32, no. 4 (2022): 2021-2051.

Rose, Jacob M. "The effects of cognitive load on decision aid users." *In Advances in Accounting Behavioral Research*, pp. 115-140. Emerald Group Publishing Limited, 2002.

Rose, Jacob M., Britton A. McKay, Carolyn Strand Norman, and Anna M. Rose. "Designing decision aids to promote the development of expertise." *Journal of Information Systems* 26, no. 1 (2012): 7-34.

Roychowdhury, Sugata. "Earnings management through real activities manipulation." *Journal of Accounting and Economics* 42.3 (2006): 335-370.

Sah, Sunita, Don A. Moore, and Robert J. MacCoun. “Cheap Talk and Credibility: The Consequences of Confidence and Accuracy on Advisor Credibility and Persuasiveness.” *Organizational Behavior and Human Decision Processes* 121, no. 2 (July 1, 2013): 246–55. https://doi.org/10.1016/j.obhdp.2013.02.001.

Schlicker, Nadine, Markus Langer, Sonja K. Ötting, Kevin Baum, Cornelius J. König, and Dieter Wallach. “What to Expect from Opening up ‘Black Boxes’? Comparing Perceptions of Justice between Human and Automated Agents.” *Computers in Human Behavior* 122 (September 1, 2021): 106837. https://doi.org/10.1016/j.chb.2021.106837.

Seong, Younho, and Ann M. Bisantz. “The Impact of Cognitive Feedback on Judgment Performance and Trust with Decision Aids.” *International Journal of Industrial Ergonomics* 38, no. 7 (July 1, 2008): 608–25. https://doi.org/10.1016/j.ergon.2008.01.007.

Shin, Donghee, and Yong Jin Park. “Role of Fairness, Accountability, and Transparency in Algorithmic Affordance.” *Computers in Human Behavior* 98 (September 1, 2019): 277–84. https://doi.org/10.1016/j.chb.2019.04.019.

Smedley, Georgia A., and Steve G. Sutton. "Explanation Provision in Knowledge-Based Systems: A Theory-Driven Approach for Knowledge Transfer Designs." *Journal of Emerging Technologies in Accounting* 1, no. 1 (2004): 41-61.

Solberg, Elizabeth, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. “A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids.” *Group & Organization Management* 47, no. 2 (April 2022): 187–222. https://doi.org/10.1177/10596011221081238.

Spain, Randall D. "The effects of automation expertise, system confidence, and image quality on trust, compliance, and performance." (Doctoral dissertation). Old Dominion University, Norfolk, VA. (2009).

Sprent, Peter and Nigel C. Smeeton. *Applied Nonparametric Statistical Methods*. CRC Press, 2007.

Steinbart, Paul John, and Wilton L. Accola. "The effects of explanation type and user involvement on learning from and satisfaction with expert systems." *Journal of Information Systems* 8, no. 1 (1994).

Sutton, Steve, Matthew Holt, and Vicky Arnold. “‘The Reports of My Death Are Greatly Exaggerated’—Artificial Intelligence Research in Accounting.” *International Journal of Accounting Information Systems* 22 (September 2016): 60–73. https://doi.org/10.1016/j.accinf.2016.07.005.

van Dongen, Kees, and Peter-Paul van Maanen. "A framework for explaining reliance on decision aids." *International Journal of Human-Computer Studies* 71, no. 4 (2013): 410-424.

Van Lent, Michael, William Fisher, and Michael Mancuso. “An explainable artificial intelligence system for small-unit tactical behavior.” In *Proceedings of the national conference on artificial intelligence* (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

Wang, Binyu, Zhe Liu, Qingbiao Li, and Amanda Prorok. "Mobile robot path planning in dynamic environments through globally guided reinforcement learning." *IEEE Robotics and Automation Letters* 5, no. 4 (2020): 6932-6939.

Wilkison, Bart D., Arthur D. Fisk, and Wendy A. Rogers. "Effects of mental model quality on collaborative system performance." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 51. No. 22. Sage CA: Los Angeles, CA: SAGE Publications, 2007.

Ye, L. Richard, and Paul E. Johnson. "The impact of explanation facilities on user acceptance of expert systems advice." MIS Quarterly (1995): 157-172.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. “Making Sense of Recommendations.” *Journal of Behavioral Decision Making* 32, no. 4 (October 2019): 403–14. https://doi.org/10.1002/bdm.2118.

Zhang, Chanyuan Abigail, Soohyun Cho, and Miklos Vasarhelyi. “Explainable Artificial Intelligence (XAI) in Auditing.” *International Journal of Accounting Information Systems*, 46 (2002): 100572.

Appendix A: Table 1A – Descriptive of Pre-Scaled Reliance Score

(Detailed Experimental Instrument)

Audit Support from Jones & Baker's SmartER System

Jones & Baker (your audit firm) has developed a proprietary artificial intelligence system (Smart Estimation of Reserves, or SmartER) that audit teams can use to help with audits of commercial loan portfolios on banking engagements. To develop SmartER, Jones & Baker partnered with a large international technology company with leading experts in artificial intelligence. Additionally, the firm gathered input from valuation specialists with expertise in commercial loan grading (e.g. advanced degrees, professional certifications, significant experience, and extensive and rigorous training). The firm has invested significant resources developing and supporting the SmartER. Using artificial neural network models, the SmartER applies firm-approved methodologies to evaluate information from from clients as well as external information to develop independent loan grades for individual loans in the client's portfolio. Each loan grade (e.g., "AAA", ... , "D") corresponds to a range of reasonable expected loss percentages that engagement teams can then use to test the Loan Loss Reserves ("LLR").

[Only appears in Reliability Conditions]
The firm has determined that **SmartER has been 96% more accurate in calculating LLR estimates than its human counterparts for the past 20 audit engagements.**
[/End Reliability]

[Only appears in Transparency Conditions]
SmartER focuses on several pivotal variables to determine an appropriate LLR for a specific client. Among these critical factors are each loan's percentage of total portfolio, past and trending interest rates, consumer debt default, changes in underlying collateral value, changes in loan volume, and severity of past-due loans. These variables are recognized by experts as influential factors for the Probability of Default ("PD") and Loss Given Default calculations ("LGD"). Both PD and LGD factor into an individual loan's grade. The LLR is then calculated from each individual loan's grade.

To illustrate, if past and trending interest rates are increasing, SmartER would calculate that as a higher loan risk. Decreasing interest rates would lower the loan risk. A higher value of a loan's percentage of the total portfolio would increase its importance, while a lower value would decrease the impact of that specific loan on the LLR. If the PD for a loan is high, then the individual loan's grade would be lower, which if the PD is low, then the individual loan grade would be higher. Likewise, if the LGD is high for a specific loan, SmartER would grade that loan lower, resulting in a higher LLR. If the LGD is low,

then SmartER would grade that loan higher, resulting in a lower LLR.
[/End Transparency]

Jones & Baker and the expert third-party consultants have performed rigorous internal testing and are confident in SmartER's ability to generate sufficient and appropriate audit evidence for the LLR. However, while SmartER aims to make well-calibrated predictions across a portfolio of loans, it will not be perfect on every loan. Though SmartER is not perfect, with exposure to new information and situations it will learn and improve over time. Additionally, the resulting estimates will reflect significant measurement uncertainty, meaning that the actual loan losses within a given portfolio could be materially different from SmartER's predictions.

Firm guidance indicates that engagement teams can use evidence from SmartER to help develop conclusions about account balances. However, because of the close interactions between audit teams and clients regarding these issues, audit teams are still free to use their own judgment.

Appendix B: Table 1A – Descriptive of Pre-Scaled Reliance Score

EXEMPT DETERMINATION

November 22, 2023
David Watson
4202 E. Fowler Avenue BSN 3403
Tampa, FL 33620

Dear David Watson:

On 11/22/2023, the IRB reviewed and approved the following protocol:

Application Type:	Initial Study
IRB ID:	STUDY006485
Review Type:	Exempt 3
Title:	Through the Looking Glass: Overcoming Algorithm Bias in Accounting
Funding:	Muma College of Business
Protocol:	Study006485_Protocol

The IRB determined that this protocol meets the criteria for exemption from IRB review.

In conducting this protocol, you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Please note, as per USF policy, once the exempt determination is made, the application is closed in BullsIRB. This does not limit your ability to conduct the research. Any proposed or anticipated change to the study design that was previously declared exempt from IRB oversight must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant a modification or new application.

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit a new request to the IRB for a determination.

Sincerely,

Gabriela Plazarte
IRB Research Compliance Administrator

Appendix C: Table 1A – Descriptive of Pre-Scaled Reliance Score

Table 1A

Descriptive Statistics of Pre-Scaled Reliance Score[a]

	Condition			
Mean (St. Dev)	**Base (n = 12)**	**Reliability (n = 18)**	**Transparency (n = 12)**	**Transparency-Reliability (n = 10)**
Proposed Adjustment	4.0 (3.4)	10.8 (7.4)	9.3 (7.0)	11.3 (6.1)
Expected Adjustment	3.0 (2.7)	7.0 (5.8)	6.4 (5.0)	7.6 (3.7)
Proposed vs Adjustment	-1.0 (1.1)	-3.8 (7.0)	-2.9 (4.2)	-3.7 (5.3)

[a] The participants' selected an adjustment value between 0 and 20 million.

Appendix D: Declaration of Use of Generative AI and AI-assisted Technologies

Statement: During the preparation of this work the author used ChatGPT 4.0 in order to improve language and readability issues in the revision process. After using ChatGPT 4.0, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication. No generative content was used and ChatGPT 4.0 was only used as an assistive tool. Full committee approval was given to use this tool for assistive work. ChatGPT is cited in the references.